DOCTOR OF PHILOSOPHY

Probabilistic Modelling of Replication Fidelity in Eukaryotic Genomes

Mamun, Mohammed Al

Award date:
2016

PROBABILISTIC MODELLING OF REPLICATION FIDELITY IN EUKARYOTIC GENOMES

by

MOHAMMED AL MAMUN
Bsc., Khulna University, Khulna, Bangladesh, 2011

THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE DEPARTMENT

OF

COMPUTATIONAL BIOLOGY

7 July 2016

UNIVERSITY OF DUNDEE

When you submit your thesis

Approval

Name	Mohammed Al Mamun
Degree	Doctor of Philosophy
Thesis Title	Probabilistic Modelling of Replication Fidelity In Eukaryotic Genomes
Examining Committee	Professor Kees Weijer (Convener)

_____
Professor Timothy Newman (Supervisor)

_____
Professor Tomo Tanaka (Internal Examiner)

_____
Professor Ian Stansfield (External Examiner)
Inst. of Medical Sciences, University of Aberdeen

Date Approved **July 2016**

Declaration

I confirm that I am the sole author of this thesis and that all references cited (unless otherwise stated) have been consulted by me. This thesis is a record of the work done by me in pursuit of the degree of PhD in Theoretical Biophysics. This work has not been previously submitted for any other higher degree.

Mohammed Al Mamun

20 July 2016

I certify that Mohammed Al Mamun has fulfilled the relevant ordinance and regulations of the University Court and is qualified to submit this thesis for the degree of Doctor of Philosophy.

Professor Timothy Newman

Professor of Theoretical Physics and Systems Biology
Vice-Principal (Research, Knowledge Exchange and Wider Impact)
University of Dundee

20 July 2016

PROLOGUE

The first scientist who invented and formalized modern scientific methodology, Ibn Al Haytham – the physicist, said:

"Truth is sought for its own sake ... Finding the truth is difficult, and the road to it is rough. For the truths are plunged in obscurity. ... God, however, has not preserved the scientist from error and has not safeguarded science from shortcomings and faults. If this had been the case, scientists would not have disagreed upon any point of science ... Therefore, the seeker after the truth is not one who studies the writings of the ancients and, following his natural disposition, puts his trust in them, but rather the one who suspects his faith in them and questions what he gathers from them, the one who submits to argument and demonstration, and not to the sayings of a human being whose nature is fraught with all kinds of imperfection and deficiency. Thus the duty of the man who investigates the writings of scientists, if learning the truth is his goal, is to make himself an enemy of all that he reads, and, applying his mind to the core and margins of its content, attack it from every side. He should also suspect himself as he performs his critical examination of it, so that he may avoid falling into either prejudice or leniency."[1,2]

Ibn Al Haytham

[1]Sabra, A. I. Ibn al-Haytham. *Harvard Magazine* (2003).
Available at: http://harvardmagazine.com/2003/09/ibn-al-haytham-html. (Accessed: 17th May 2016)
[2]Ibn Al Haytham, *Doubts Concerning Ptolemy*, Translated by S. Pines, as quoted in Sambursky 1974, p. 139.

Abstract

Eukaryotic DNA replication is composed of a complex array of molecular biological activities compounded by the pressure for faithful replication in order to maintain genetic and genomic integrity. The constraints governing DNA replication biology is of fundamental importance to understand the degree of replication error and strategies employed by organisms to tackle the threats to replication fidelity from such errors. We apply a simple conceptual model, formalized by the use of probability theory and statistics, to discern fundamental pressures and constraints that optimise complete DNA replication in genomes of different size scales (10 Megabases to 10 Gigabases), spanning the whole *eukaryota*.

We show in yeasts (genome size ~10 Megabases) that the replication origins (sites on DNA where replication can be initiated) are biased towards equal spacing on the genome and the largest gap between adjacent origins is limited compared to that is expected by chance, as well as origins are placed very close to the telomeric ends in order to minimize the replication errors arising from occasional irreversible failures of replication forks. Replication origin mapping data from five different yeasts confirm to all of these predictions. We derive an estimate of $\sim 5.8 \times 10^{-8}$ for the fork stalling rate per nucleotide, the one unknown parameter in our theory, which conforms to previous experimental estimates.

We show in higher eukaryotes (genome size 100 Megabases to 10 Gigabases) that the bias for equal origin spacing is absent, larger origin gaps contribute more to the errors

while the permissible origin separations are restricted by the rate of fork stalling per nucleotide, and in the larger genomes (> 100 Megabases) errors become increasingly inevitable, yet with low net number of events, that follows a Poisson with small mean. We show, in very large genomes e.g. human genome, that larger gaps contributing most to the error are distributed as a power law to spread the risk of damage from the error, and post-replicative error-correction mechanisms are necessary for containment of the inevitable errors. Replication origin mapping data from yeast, *Arabidopsis, Drosophila* and human cell lines as well as experimental observations of post replicative error markers validate these predictions.

We show that replication errors can be quantified from the nucleosome scale minimum inter-origin distance permissible under the known DNA structure and we propose a universal replication constant maintained across all eukaryotes independent of their architectural complexity. We show this molecular biological constant relates the genome length and developmental robustness of organisms and this is confirmed by early embryonic mortality rates from different organisms.

Good agreement of the biologically obtained data to the model predictions in all cases suggests our model efficiently captures the biological complexity involved in containing errors in the DNA replication process. Conceptually, the model thus portrays how simple ideas can help complex biology to elevate our understanding of the continuously increasing knowledge of biological details.

Dedication

I dedicate this work to all the innocent children dying from unjust wars, oppressions and tyranny from Syria to Burma and throughout the world. They remind me the purpose of life and death, and inspire me to do things better than I am doing.

Acknowledgements

'*O Allah, You are the First and there is nothing before You; You are the Last and there is nothing after You. You are the Manifest and there is nothing above You; You are the Hidden and there is nothing beyond You.*' I praise and glorify You with utmost gratitude and humbleness for all the blessings and richness of this world that You have bestowed upon me; for all the mercy and grace that You have showered upon me and for all the happiness and love that You have poured around me. Indeed, You have granted me richness after poverty and knowledge after ignorance. Indeed, I am grateful to You for allowing me to ponder and to reflect upon some of Your creations using reason and rational that You have gifted to the mankind as a way to discern the knowledge of truth from the ignorance of doubts and superstitions. Indeed, the first scientist Ibn Al Haytham rightfully said, '*For closeness to God, there is no better way than that of searching for truth and knowledge*.' Indeed, I thank You for making me one of the seeker of truth in the faculty of reason and for this I use the methodology of science to study the biology of living things. I do this for Your sake only and grant me Your closeness by the virtue of whatever good in it and protect me from whatever evil in it. Indeed, we do not have any knowledge except by Your permission, so allow us knowledge that is beneficial for the humanity from this humble endeavour.

I could not be me without them. So, I remember them and thank them from the core of my heart – they are my parents. They have done what they have done and I know how much they sacrificed for our school fees. May Allah reward them abundantly for all their supports and sacrifices. I thank my two brothers and the only sister and obviously thanks to my wife who supported me greatly during the last year of my PhD.

This thesis would not be a fact except my supervisor Professor Newman. I am from a culture where the teacher is seen as an extended family guardian and Indeed, I found professor Newman like that with his humble etiquettes, cordial attitudes, personal concerns and supporting advices ranging all over my work and life throughout these

years as well as for the future. Whatever scientific achievement we had together is from his tea time pencil drawings and obviously this has taught me greatly how simple reasoning can help understand very complex problems which is actually not only a scientific gain but also a valuable learning for practicality surrounding us.

I am deeply thankful to Luca Albergante whose help and support during whole of my PhD is more than my thankfulness. Whatever little programming and statistics I learned, I learned it from him and moreover his contributions to my projects were so much so that sometime I felt he was doing for me more than me and I can not be more thankful for the excellent collaboration we had. I am also very much grateful to Dianbo Liu and Sam Palmer. Numerous group discussions we had all together regarding boundless topics in different sciences had it's own teachings for me as well as contributed to my projects tremendously. I thank Professor Julian Blow, Alberto Moreno and James Carrington not only for the great collaboration and impressive experiments that I had the opportunity to use for the validation of our theory but also for generous comments and discussions regarding the theory to make it more biologically fruitful. And at last but not the least is SLS, University of Dundee and SULSA; I acknowledge their support in funding and sponsoring my PhD.

"Indeed, in the creation of the heavens and the earth, and the alternation of the night and the day, and the ships which sail through the sea with that which benefits people, and what Allah has sent down from the heavens of rain, giving life thereby to the earth after its lifelessness and dispersing therein every [kind of] moving creature, and [His] directing of the winds and the clouds controlled between the heaven and earth are signs for a people who use reason."[1] When they reflect on the grandeur of nature, they are deeply moved and exclaim*: "Our Lord! (Rabb) Thou hast not created this in vain."*[2]

[1]Quran, Surah Al-Baqarah, **2**:164
[2]Quran, Surah Ali 'Imran, **3**:191

List of Tables

List of Figures

Contents

Chapter 1
Introduction

1.1 Probability theory in biology: an overview

Probability theory defines the likelihood of a specific result in an experiment/event. For example, what is the probability of a coin to show heads in a free toss or what is the likelihood of it snowing in Scotland tomorrow, are issues discussed in probability theory. In present days, the public is well accustomed to the phrases like “smoking increases the probability of getting cancer by thirty percent” or “regular exercise decreases the probability of diabetes”. More science specific examples are what is the probability of a particular gene to be engaged in oncogenesis or what is the probability of a certain trait to be linked to a particular set of genes. These are common examples of how probability theory is applied from the public to the specialists in biological background.

The role of probability in understanding biology is multi-scale and multi-dimensional. It begins from the very basic constituent of matter itself i.e. the sub-atomic world. The probabilistic description of the wave function in ‘Quantum mechanics’ was a gigantic breakthrough in understanding the emerging outcome from the ‘spooky’ actions in the sub-atomic realm (Dirac, 1982; Shalm et al., 2015). Thus probability is fundamental in understanding the transition from the inherently uncertain quantum world to the chaotic dynamics in the sub-cellular bio-molecular world. DNA -> mRNA -> protein; is the simplest causal diagram that represents the huge world of bio-molecules. Even though at single gene level, a gene to its protein product follows deterministic behavior but at a systemic level, interplay between genome and proteome coupled with the

phenotypic outcomes shows large scale indeterminacy and probability based network dynamic modeling strategies offer a great potential in helping to understand this sub-cellular element world. Stochastic noise induced fluctuations and molecular chaos inside the cell contributes to the cell biology and biological mechanisms at cellular scale and subsequent organic and organismic behavior, i.e. survival and homeostasis. Living organisms are in constant interaction with their surrounding environment. Simultaneous internal and external continuum of different forces and factors are responsible for adaptation and survival of an organism in a particular niche i.e. cells in microenvironment, organisms in ecosystem. The survival here would be the increase in probability of favorable outcomes in relevant biological reactions (Nakajima, 2013). Biological homeostasis or a disease condition are related to increase or decrease in such probabilities.

In the last decade or so, the biological community has gone through a huge revolution in data production. Big data on molecular details of different biological mechanisms have made it possible to look at biology from a holistic perspective – how biology works? Random molecular interactions and fluctuations e.g. protein-protein interaction or gene transcription, which are responsible for one scale up from cell biology like cell division or DNA replication, can be statistically simplified in order to ask bigger questions as such. The implication of inherent stochasticity and randomness of molecular biology needs to be understood systemically in the context of complete and faithful DNA replication, successful cell division or stable stem cell pool, which are fundamental for the active biological systems. Thus mere use of probability to describe the chance of getting cancer from smoking, needs a phenomenological shift in the way it is used in biological context to better learn why biology is as it is. In this thesis, we present an example of such an initiative. We describe a model based on probability

theory for DNA replication that explains basic constraints for replication fidelity at different scales across eukaryotes and connects molecular level deterministic biology, i.e. the genome and its organization to the higher order biology of replication and development.

1.2 Simple modeling: an overview

A model is constituted based on an assumption or a set of assumptions with few (one or more) parameters that can be adjusted, estimated or experimentally defined. The assumptions are formalized in the form of mathematical equations or computational programs. Complicated assumptions can produce very complex models, which might have multiple adjustable parameters, but simplicity is compromised. The modeler has to deal with a trade-off between simplicity and fit for the purpose (Forster, 2001). Simplicity of a model can be gauged from the ratio of model outputs to its inputs where inputs represent the assumptions upon which the model is based and outputs are the resulting novel concepts and predictions from the model. Small number of inputs and greater number of outputs would imply a stronger model and higher the ratio of output to input, simpler is the model; alternatively lesser ratio of output to input could often lead to weakening of the predictive power of the model as well as parameters can become less precise (Newman, 2015). In terms of practical effectiveness, simple assumptions can produce powerful inferences and predictions that are essential attributes of a powerful model. Hence, we pursue simple modeling. In this endeavor, we have a simple methodology. Primary formalization (computational or mathematical) of basic idea or assumption regarding the problem in hand is to be continuously confronted with experiment and data. The primary assumption either survives the challenge and yields auxiliary inferences, or because of the inherent simplicity it can be easily shifted to new assumptions; which gives a broader scope for

the model to face logical inductions from a scientific view point. Thus, the model becomes a powerful tool to ask and investigate questions on a broader systemic level.

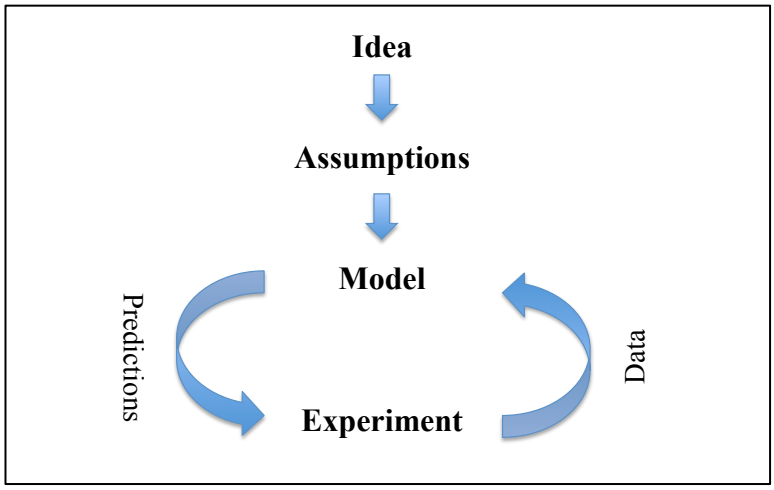


Figure 1: Schematic of the methodology we use in simple modeling.

In physical sciences there are examples of very simple models successfully explaining very complex phenomena, for example Boltzman's assumption of equal a priori probabilities underpins the whole of statistical mechanics. Likewise, a simple model with strong inferences, like the Volterra predator-prey model in ecology has had a long and effective usefulness. Inspired from these examples, the group I am working in has recently looked at problems like DNA replication, metastasis in cancer, and gene regulation, from a simpler point of view. A rare event theory of cancer metastasis (Cisneros and Newman, 2014), Buffered Qualitative Stability of gene regulatory networks (Albergante et al., 2014), are successful examples for the group's view of simple modeling. Under the same motivation this thesis is also another example for such simple modeling strategy. Our model here is based on a very simple probabilistic assumption that each nucleotide replicated by the replication fork has an equal minute probability for the fork to stall irreversibly. Even though the relevant biology is very complex, discussed in detail in the following section, yet with this simple assumption

the model captures important systemic features of the complex biology as will show in the following chapters.

1.3 DNA replication: an overview

After discovering the DNA double-helix model in 1953, Watson and Crick proposed the hypothesis of 'semiconservative replication' (Watson and Crick, 1953). Later in 1958 the hypothesis was confirmed experimentally in *Escherichia coli* (Meselson and Stahl, 1958) following the ground breaking discovery of DNA polymerase in 1956 (Kornberg et al., 1956). During replication each of the DNA strands act as template for naive strand to form and specific pairwise bonding between nucleotide bases, adenosine (A) to thymine (T), cytosine (C) to guanine (G), confirms the exact complementarity of the mother and daughter strand. In order to replicate, the double-helix needs to be unzipped by DNA helicases which break the bond between base-pairs. The positions on DNA where this unzipping starts are known as replication origins (ROs). In *E. coli* there is only one RO but in eukaryotes they range from hundreds to thousands. As soon as the DNA is unzipped, DNA primase binds the small RNA primers to DNA followed by DNA polymerase elongating the initial primer, which ultimately becomes the daughter strand. In completion, two complementary pairs of DNA double-helix is formed. This whole process can be coarse grained as RO licensing, initiation of replication, elongation and coalescence.

RO licensing is the process that determines the potential ROs across the DNA by recruiting the Origin Recognition Complex (ORC). ORC with the help of Cdc6 and Cdt1 proteins recruits the helicase Minichromose Maintenance (MCM) 2-7 hexamer. RO licensing is restricted to the end of G1 of eukaryotic cell cycle, as soon as initiation of replication kicks off in S phase, licensing machinery is inactivated providing the cell

with a safeguard from re-replication of copied DNA (Blow and Dutta, 2005; Blow and Hodgson, 2002; Remus et al., 2009).

In S phase MCM2-7 hexamers are activated following the recruitment of Cdc45 and GINS, which together acts as the replicative helicase that unzips the DNA, moving bidirectionally along the double-strand. These bidirectional unzipped domains of DNA caused by the helicase complex are known as replication forks (RF) (Riera et al., 2014; Vijayraghavan and Schwacha, 2012).

DNA polymerase is recruited following the unzipping of double-strand by the helicase. Polymerase does the synthesis of nucleotide bases only from (5′-3′) direction thus the leading strand (3′-5′) is replicated continuously as the complementary daughter strand is in opposite direction. (5′-3′) lagging strand is replicated discontinuously in small fragments called ‘Okazaki fragments’ which are later joined together by DNA ligase enzyme (Burgers, 2009; Zheng and Shen, 2011).

Termination of typical replication occurs when two RFs from opposite direction meet. This coalescence of RFs can potentially happen at any place between adjacent ROs. If a fork fails then the other coming from the opposite would complete the replication of remaining region of DNA and collide with its counter-part. Upon the collision, helicase complex is disassembled and DNA ligase joins the daughter strands (Leman and Noguchi, 2013).

Faithful replication of the complete DNA is essential for the genetic material to be carried across generations under a robust safeguard that would confirm the survival of the species itself over time. Eukaryotic cells activate hundreds to thousands of ROs during S phase to accomplish this fundamental task (Alver et al., 2014). As we

discussed earlier, each RO gives rise to a bidirectional pair of RFs, which essentially moves through the DNA double-strand and paves the way for DNA polymerase to work. These RFs are highly reliable even though rarely they might collapse irreversibly for different reasons. For example, damaged DNA could cause the fork to stall but more specific reasons for irreversible RF stalling is an active field for investigation (Cobb et al., 2005; De Piccoli et al., 2012). We already mentioned the biological constraint for inactivation of RO licensing before replication initiates. So, during replication cells cannot license new ROs in order to compensate for the RF failure as that could effectively open the door for re-replication of replicated DNA, which is fatal for the genetic integrity. Thus the cell must ensure that enough ROs have been licensed before the end of G1 and in order for this reason eukaryotic cells license many more ROs than essential to finish the whole genome duplication. Hence many of the licensed ROs remain unused and are passively replicated by the active RF, because the active RF causes the unfired MCMs to fall off the chromosome. These unused ROs are called dormant origins in the literature and dormant origins are indeed 3-10 fold higher in number than the activated ROs (Blow et al., 2011; McIntosh and Blow, 2012). Due to inefficient loading of MCM2-7 double-hexamer, origins could become relatively incompetent and may fail to fire in majority of cell cycles (Evrin et al., 2009). Nevertheless, the number of licensed ROs in the cell along with their distribution over the genome seems to be more directly responsible for complete replication of the genome rather than the efficiency of individual ROs (Blow and Ge, 2009). Hence, dormant origins provide a reserve contingent for emergency in the face of collapsed or stalled RFs.

The genomic location of ROs in eukaryotes is a vast field of research. In simple eukaryotes e.g. *Saccharomyces cerevisiae*, position of ROs are determined by specific

genomic sequence while in higher eukaryotes it is not so straightforward, and no substantial sequence specificity is observed (Leonard and Méchali, 2013). However the distribution of the licensed RO positions is categorical in maintaining replication fidelity (Aparicio, 2013). Within a genomic region bound by two adjacent ROs, coalescence of two RFs coming from opposite directions marks the termination of replication. These forks could either be from the two adjacent ROs (if both are activated) or could be from distant ROs that have travelled through the inactive ROs. The important issue here is within the region bounded by adjacent ROs, if one RF fails then the other RF will travel up to the failed one to complete the replication of the remaining DNA. The worst case would be both the RFs fail before they meet each other with no dormant RO in between. In this scenario the piece of DNA in between the stalled RFs would remain unreplicated and this could cause severe consequences towards genomic integrity. This potential situation with both RFs stalled inside the genomic region bounded by two adjacent ROs, is called a double fork stall (DFS). Similar unreplicated DNA could arise at the end of the linear eukaryotic chromosomes where the very end namely telomeric end is replicated by a single RF either arising from activated end-proximal RO or from distant RO which has travelled through the end-proximal RO. Irreversible stall for this lone RF, before it completes the replication of remaining DNA would result in similar consequence as for DFS. We call this terminal lone stall as telomeric fork stall (TFS). Thus, these DFSs and TFSs are tremendously significant impediments that cells must overcome in order to ensure complete replication. DNA double stranded breaks (DSBs), which is a major DNA damaging event has been linked to DFSs (Curtin and Sharma, 2015; Unno et al., 2013). Different pathologies including cancer have been associated to DFS induced aberrant DNA structures and mismanagement of such abnormalities (Abbas et al., 2013; Ghosal

and Chen, 2013; Mazouzi et al., 2014).

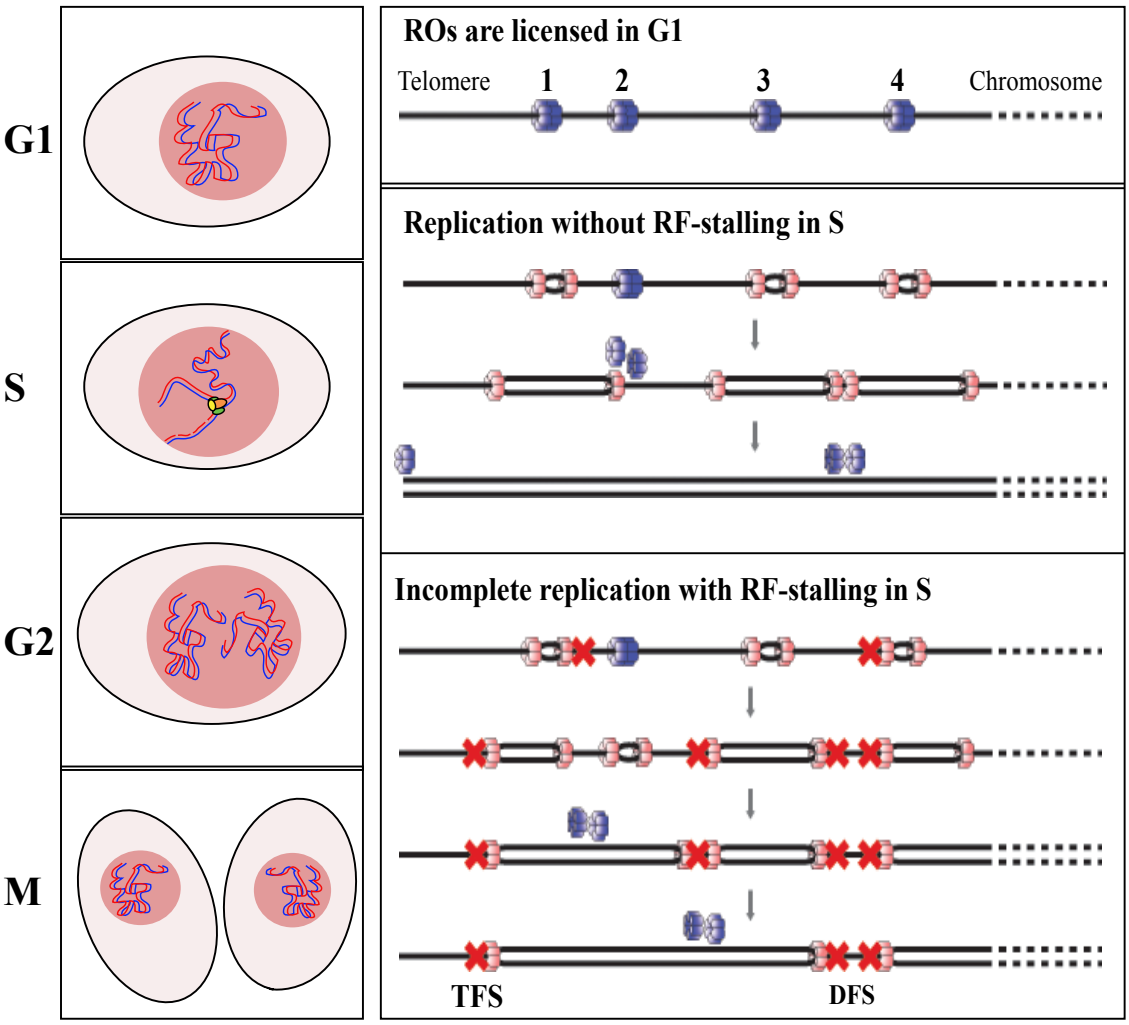


Figure 2: Different stages of eukaryotic cell cycle G1, S, G2 and M with the genome duplication status for each cycle inside the nucleus is shown on the left. On the right, RO licensing in G1, complete replication in S in absence of any RF stalling and incomplete replication due to RF stalling, DFS and TFS is shown schematically (Newman et al., 2013).

In figure 2, replication status in different cell cycle and in absence and presence of DFS and TFS are presented. In G1, ROs are licensed through the successful loading of MCM2-7 double hexamer. 1, 2, 3 and 4 are graphic ROs, which are licensed. In following S phase, ROs 1, 3 and 4 are fired/activated and bidirectional RFs emanated from each of them which travels through the DNA while polymerase replicates. RO 2

is passively replicated (being discarded from the DNA) by the active RF coming from RO 1. Coalescence of the pair of RFs travelling towards each other in each segment of DNA bounded by adjacent ROs completes the replication process. In case of incomplete replication, due to the stalling of the RF from RO 1 moving towards the chromosome, RO 2 is activated to complete the replication but both RFs facing each other from RO 3 and RO 4 stalled (DFS) keeping the intervening DNA unreplicated. Similar unreplicated DNA segment produced when the lone RF from RO 1, going towards the end, stalled before completing replication of remaining DNA at the telomere (TFS).

A huge amount of work has been done on studying the location of ROs in eukaryotes ranging from yeasts to humans and there is significant success in yeasts in obtaining genome wide distribution of ROs (Siow et al., 2012). However only recently genome wide distributions are obtained in metazoan cell lines i.e. different human cell lines (Besnard et al., 2012; Picard et al., 2014). Still the factors and constraints that help to establish the number and distribution of licensed ROs in eukaryotic cells, fundamental for complete and faithful genome duplication as we have discussed already, remains elusive, but the genome wide positions of ROs in eukaryotes are a huge step forward in this regard. In this thesis, we focus on one such major issue i.e. DFS and TFS (as per availability of data) arising from irreversible fork stalls and how they contribute in determining the required number and distribution of licensed ROs to manage these errors; and more in effect we study how replication fidelity is maintained in different eukaryotes from yeasts to metazoans in the face of these impediments.

1.4 Modeling DNA replication: an overview

Eukaryotic genomes require complete replication only once every cell cycle. Total biological process involved in this genomic duplication process is very complex and modelling can be done at multiple scales i.e. biochemical and structural studies of the machineries involved at molecular level, organisation and management of the replication process at a systemic level.

Mathematical modelling has been used to investigate the dynamics of the molecular network responsible for RO licensing, and to investigate the network dynamics in terms of preventing re-replication of genome segments in *Saccharomyces cerevisiae* (Brümmer et al., 2010), which is an example of a model relating both the molecular and systemic scale of DNA replication. This model suggested a trade-off between controlling the re-replication and dynamics of RO firing as rapid and synchronous firing of ROs could increase the probability of double replication too. Hence, differential timing profile of RO initiation i.e. early and late firing ROs in *S. cerevisiae* could be the result of the requirement for robust containment of the probability of re-replication of any DNA segment which is the other extreme of under replication due to DFS or TFS errors.

Earlier, RO licensing and activation have been modelled as Poisson processes and different statistics e.g. genome replication time, Okazaki fragments and number of active ROs in S phase were studied using the Poisson model (Cowan, 2003). In this applied probabilistic modelling, ROs were considered to be Poisson distributed over the long DNA molecule through an exponentially distributed timing for licensing factors to bind the specific RO sites. The proportion of licensed ROs to be activated was calculated to be roughly ~15% which was an early indication for the now known

abundant dormant ROs on the genome which is actually 3-10 fold higher than the number of activated ROs in a particular cell cycle.

Coarse grained molecular detail has been connected to the replication kinetics in an analytical model with probabilistic licensing and stochastic initiation of ROs in S phase as well as to study evolvable RF dynamics in inhomogeneous DNA replication based on the stochastic nucleation and growth model for first order phase transition in statistical physics. RF creation from firing of ROs, propagation of RF through the DNA and the coalescence of RFs was modelled based on differential rate equations for RF population in both time since S phase started and spatially the length of genome replicated and genomewide RF density was determined from this evolving kinetics of RF population. (Gauthier et al., 2012; Yang et al., 2009). The contrast of this modelling approach with ours is that they studied RF stalling at DNA damage sites while we assumed individual nucleotide level tiny probability of RF stalling as a constant.

Stochastic models coupling spatial dynamics such as location of origins along with the temporal features such as RO initiation and RF progression has been used to study the evolving replication kinetics as S phase progresses (Lygeros et al., 2008). This model showed randomly generated large inter-RO gaps would increase the overall replication timing in fission yeast, which is in very good agreement to the results of our study in yeasts. More recently RO locations and corresponding spatial dynamics for completing whole genome duplication has grabbed more attention as only recently genome-wide RO positions are revealed for more sophisticated eukaryotes as more and more evidence is piling up to support the narrative that spatial organisation of ROs plays a

major role in determining differential replication timing in eukaryotes (Aparicio, 2013).

Random fluctuations in RO licensing and stochasticity in activation of ROs drives the variability in replication timing. RO positions have been modelled in accordance to the minimum time required to complete genomic replication (Karschau et al., 2012). This model reported that the strict timing profile for replication of genomic DNA could lead to the clustering of ROs as had been observed in different eukaryotes. Also, less competent ROs in yeast can be grouped together to complete the replication process without time delay, which has a sharp contrast to our observation of RO distribution in yeasts are biased for regularity and this observed regularity is not related to the competence of individual ROs rather the actual number of licensed ROs is the main factor with their location on the genome which could be in effect related to the replication timing.

All these models including mathematical and computational, analytical and numerical, stochastic and deterministic approaches have shed light on the different scales of DNA replication biology over time. Still an integrated holistic model is lacking that can pin down the global constraints responsible for faithful replication of the complete genome during S phage of eukaryotic cell cycle. In the current thesis, we have considered all licensed ROs to be potentially active and exempts the replication process from a strict timing issue that could be found in embryos. Thus our model is more reflective of the somatic cells in adult organisms, which is generally exempted from the strict timing factor. Giving the chance of a RO to be inactivated by the actively travelling RF, we do not take into account the individual competency of ROs rather we focus on individual nucleotides being replicated. Our model begins with a simple basic assumption that

each nucleotide has an equal minute probability for the travelling RF to stall irreversibly. Fork stalling has been linked to gross chromosomal instability in an oncogenic model for development of cancer (Aguilera and Gómez-González, 2008; Halazonetis et al., 2008). Also density of stalled forks have been proposed as indicator for a cell to be normal or cancerous (Gauthier et al., 2010). In this light we asked what is the chance of a cell to duplicate its genome faithfully in the face of irreversible fork stalling events i.e. DFSs and TFSs, discussed in the previous section. Using a probabilistic model, we quantify the chance of replication failure from such events and compare this to the experimentally observed replication error rate in different organisms. We predict the organisms with smaller genomes e.g. yeasts manages the negligible error rate by maintaining a regular RO distribution while larger genome lengths requires error correction mechanisms due to the emerging inevitability of replication errors. In our model, we also relate the genome scale replication error rate and developmental robustness of organisms to the measurable molecular factors i.e. nucleosome distance and per nucleotide fork stalling rate. In all cases we compare the model predictions to the experimentally observed data as a test and potential validation of the model.

1.5 Structure of this thesis

The thesis is structured as follows: we first construct the model and show the mathematical derivations in chapter 2. Following that, we apply the model to yeasts and analyze RO positions genome wide in different yeasts in chapter 3. In chapter 4, we apply the model to higher eukaryotes and use different whole genome RO position datasets including yeasts, *Drosophila*, *Arabidopsis* and human for the analysis. In chapter 5, we apply the model to relate molecular biologically conserved nucleosome

distance and per nucleotide fork stalling rate to the genome length and early developmental robustness. We analyze data from different organisms to test the predicted relation. Lastly, we present concluding notes and highlight the future directions in chapter 6. The whole thesis is based on the theme of simple modeling and perspective over view of the relevant topic is presented in the beginning of each chapter.

Chapter 2

The Model

2.1 Mathematical introduction: an overview

My supervisor Professor Newman initially constructed the primary platform for the mathematical model. The present form of the mathematical model was developed from there and the calculations were carried out mostly during numerous tea-time discussions we had together regarding the relevant biology and data analysis that I was doing in my times. In order to construct the model, we use various mathematical and statistical definitions, laws and theorems. Before going into the model description, we first provide a general introduction to these concepts here in the beginning of this chapter.

Mean, median, variance and standard deviation

In probability and statistics, mean is the measure for central tendency that defines the arithmetic average of a set of random variables characterized by a probability distribution (Yates and Goodman, 2004). Formally, the summation of all possible value k for K random variables multiplied by its probability $P(k)$ will be called as population mean or the expected value for the variable, denoted by μ. Hence,

$$\mu = \sum k \; P(k).$$

Mean can also be defined for a set of sampled values k_1 to k_n, given a sample size of n, for K random variable which is called as the sample mean $\overline{K_n}$ and is given by

$$\overline{K_n} = \frac{1}{n} \; (K_1 + K_2 + \; K_3 + \; \ldots + \; K_n).$$

Median defines the point or number in a set of random variables, which separates one half of the data from other half. Breakdown point at 50% makes the median a very resistant statistic and hence it is very important in robust statistics. In a probability distribution *P* for *K* random variable, the real number 'm' will be the median if

$$P(K \geq m) = \frac{1}{2} \text{ and } P(K \leq m) = \frac{1}{2}.$$

Variance of a set of random variables defines the distance of how far the values are distributed from the mean. Variance is non-negative and small variance means the data points are not far from the mean and are close to each other while large variance suggests that the data points are widely spread away from the mean and also are widely apart from each other. Formally, the variance in a set of random variables or data points that is described by *K,* is the second central moment of *K*. It is given by the average or expected value of squared deviation from the mean μ of the variable *K*.

$$Var(K) = \overline{(K-\mu)^2}.$$

Standard deviation also like variance measures the dispersion of a set of data points or random variables but it is the square root of the variance of the data or variable. For K random variable that has the mean μ, standard deviation (*sd*) is given by

$$sd(K) = \sqrt{\overline{(K-\mu)^2}}.$$

Coefficient of variation

In statistics the ratio of standard deviation to mean is also known as 'coefficient of variation', which we denote here with *R*. For points on a string, *R* is the ratio between standard deviation (sd) and mean (μ) for the distances between adjacent pair of points (Everitt, 1998).

$$R = \frac{sd}{\mu}.$$

Periodic spatial ordering of some points on a one-dimensional string would provide $R = 0$ and any deviation from such periodicity would give a non-zero positive value to R. Random distribution of those points would essentially establish a contextual upper bound depending on the number of points considered. Under simulation we checked this tendency of upper bound for the R value in random distribution and different number of points sampled randomly on a finite line provides an upper bound for R around 1.

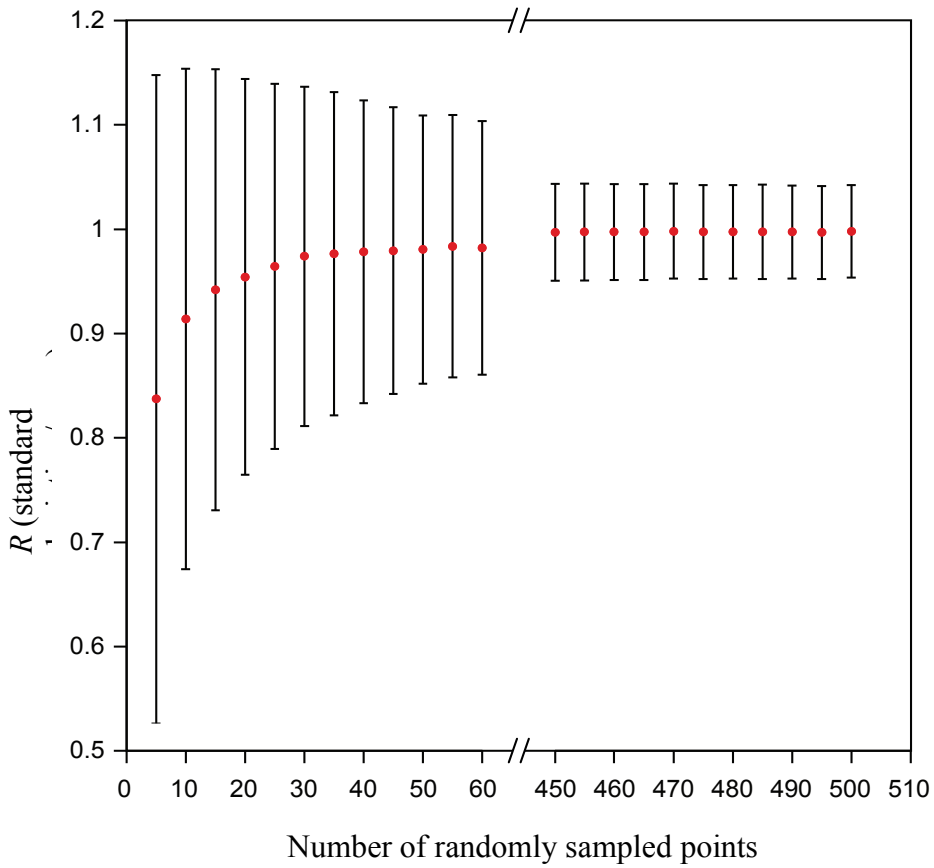


Figure 3: R for randomly sampled points on a finite line. Points were randomly sampled with 10,000 iterations and the mean of the value of R is shown in red dots with error bars showing the respective standard deviation.

Any R bigger than 1 would imply complex form of distribution for the distances between sampled points far away from periodicity. Below is a schematic diagram with

20 points sampled on a line in such way that they give $R = 0$ in the first instance, $R = 1$ in the second case and $R = 2$ in the third scenario.

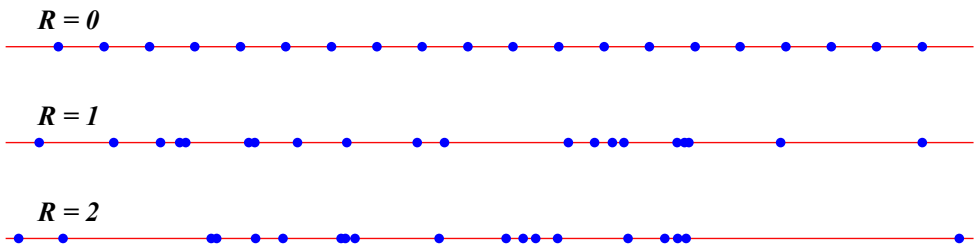


Figure 4: Schematic of the different values of coefficient of variation, R, in the distribution of points on a straight line.

Essentially, R is an intuitive measure for the degree of regularity in a distribution of points on a string or line: R = 0, the points are in periodic order, if R ≈ 1, they are randomly distributed and if R > 1, the distribution is far away from periodicity or randomness.

Law of large numbers

In probability and statistics, law of large numbers is a principle that describes the mean in a large number of sampling in a random process is close to the actual expected value of the process. According to the law, as the number of samples increases, the percentage difference between sample mean and expected value goes towards zero (Geiringer, 1940). Let, $K_1, K_2, K_3, \ldots, K_n$ be independent samplings in a random process with a finite expected value of $\varepsilon = E(K_i)$. According to the law of large numbers, the sample mean

$$\overline{K_n} = \frac{1}{n}\,(K_1 + K_2 + K_3 + \ldots + K_n),$$

converges as

$$\overline{K_n} \longrightarrow \varepsilon, \text{ when } n \longrightarrow \infty.$$

Poisson distribution

In probability and statistics, the Poisson distribution is a discrete probability distribution used to express the probability of a number of mutually independent events where the rate for the occurrence of the event is constant over unit time or space. How many events will occur within a unit time or space in a system with multiple possible events, where each of the events is independent from the other events, is a random variable with Poisson distribution (Haight, 1967). Formally, a discrete random variable *X* will be called a Poisson distribution with non-zero positive parameter 'λ' if the probability of 'k = 0, 1, 2, …' occurrences is given by the following probability distribution function:

$$P(X = k \mid \lambda) = \frac{\left(e^{-\lambda}\right) \lambda^{k}}{k!},$$

Where *e* is the base for natural logarithm, *e* = 2.71828…, *k* is the number occurrences of the event and parameter λ is the expected number of occurrences in the given unit of time or space which is a positive real number (Yates and Goodman, 2004). For Poisson distribution, only parameter λ defines the expected number and mean as well as variance of the distribution. If we calculate the Poisson probability distribution function for zero events, $k = 0$, we get:

$$P(\text{zero events}) = e^{-\lambda}.$$

2.2 General Model

We first construct a general model that describes the probability of a DFS (double fork stall) within a region of DNA bounded by two adjacent pair of ROs (replication origins). Later we will expand this general model in accordance with the biological context of model application in the following sections.

Assumptions and definitions used in the model

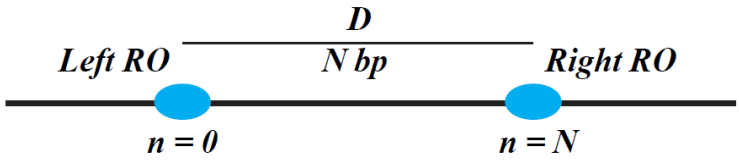
a) RFs can only originate at licensed ROs and licensed ROs are established at specific sites on the genome prior to the initiation of replication process,
b) bidirectional RFs are activated as a licensed RO is fired and thus the RO becomes unlicensed and the licensed ROs yet to fire can be passively replicated by a travelling RF originated from another RO,
c) each RF has a constant independent probability 'q' per nucleotide for irreversibly stalling (or otherwise failing),
d) any given inter-RO distance is designated by N_i and the average separation (in base pairs) between licensed ROs over the given genome is defined to be N_l,
e) the total length of the genome is defined to be N_g,
f) the median stalling distance of a RF is defined to be N_s,
g) DNA segment at the extreme ends of a chromosome that extends from the last RO (the 'subtelomeric RO') to the telomere represents a special case as it can only be replicated by a single RF and we will take this into consideration according to the data availability for the positions of such 'subtelomeric RO',
h) we consider no strict time constraint for replication of the whole genome.

We have summarized these baseline assumptions of the model in the schematic diagram presented in Figure 2 in previous chapter that represents the profile of DNA

replication in presence of DFS error as well as the complete replication in absence of such error. Based on these assumptions and definitions we proceed to our model. First we will calculate the general form for the probability of DFS within a region of DNA bounded by two adjacent licensed ROs. Later we will use this general expression for small and large genomes spanning Megabase to Gigabase across eukaryotes accordingly with appropriate conditions.

Probability of DFS between two adjacent ROs

We define by *D*, the segment of DNA between two adjacent ROs, and denote the nucleotide bases in *D* by an integer variable *n*. Let the left RO be located at $n = 0$, and the right RO be located at $n = N$.



The probability of a 'double fork stall' (DFS) in *D* is calculated by the following expression:

$$P(\text{DFS in } D) = \sum_{n=0}^{N-1} P(\text{stall from left at position } n < N) \times P(\text{stall from right at any position } n' > n) \qquad (1).$$

Note that this same expression is true even if the left RO and right RO initiate replication at different times. Whether the adjacent ROs fire simultaneously or not, this is a question of time dependent details and we simplify this by considering no time restriction on the replication process. This allows us to calculate the probability of double fork stalls between adjacent ROs considering the progression of RF by individual nucleotides and hence as long

as one RF replicates the DNA bases and stalls on a nucleotide and same happens to the other RF coming from opposite direction before two RFs meet each other within a region bounded by two adjacent ROs, our Eq. (1) is valid. Similarly this is true even if one or both the adjacent ROs are passively replicated by active RFs.

Now, since 'q' is the mean per-nucleotide stall rate:

$$P(\text{stall from left at position } n < N) = (1 - q)^n q \qquad (2).$$

Similarly,

$$P(\text{stall from right at position } n' > n) = (1 - q)^{N-n'} q \qquad (3).$$

We have to sum Eq. (3) over all possible $n' > n$, in order to get the total probability of a stall of the RF (i.e. left-moving) from the right RO that occurs at a site to the right of the stalled RF from the left RO located at n. So,

$$\begin{aligned} P(\text{stall from right at any position } n' > n) &= \sum_{n'=n+1}^{N} (1 - q)^{N-n'} q \\ &= \sum_{n''=0}^{N-n-1} (1 - q)^{n''} q \\ &= 1 - (1 - q)^{N-n} \qquad (4). \end{aligned}$$

For clarity, we have defined a new summation variable $n'' = N - n'$ for the sum, and used the following formula for summation of a geometric series:

$$\sum_{m=0}^{M-1} r^m = \frac{1 - r^M}{1 - r} \qquad (5).$$

Inserting together Eqs. (2) and (4) we have

$$P(\text{DFS in } D) = \sum_{n=0}^{N-1} (1 - q)^n q [1 - (1 - q)^{N-n}]$$

$$= q \sum_{n=0}^{N-1} (1-q)^n - q \sum_{n=0}^{N-1} (1-q)^n (1-q)^{N-n}$$

$$= q \sum_{n=0}^{N-1} (1-q)^n - q \sum_{n=0}^{N-1} (1-q)^N$$

Evaluating the sums using the Eq. (5) we have,

$$P(\text{DFS in } D) = q \left(\frac{1-(1-q)^N}{1-(1-q)} \right) - Nq\,(1-q)^N$$

$$= 1-(1-q)^N - Nq\,(1-q)^N$$

Thus,

$$P(\text{DFS in } D) = 1-(1+Nq)(1-q)^N \qquad (6).$$

Eq. (6) represents a general expression for the probability of DFS errors in any given region of DNA in between two adjacent licensed ROs. In the following sections, we will use this general model appropriately for derivations of specific formulas for genome-wide probability of replication errors according to the relevant biological context of different genome lengths and inter-RO gaps. As our aim is to constrain all the mathematics in this one chapter, the different derivations and formulas throughout section 1.3-1.5 will be used and contextualized in rest of the thesis accordingly.

2.3 Model A

For small genomes of around 10 Mbp with hundreds of ROs spread out over the genome as found in simpler eukaryotes like yeasts, we assume the hierarchy: $N_s \gg N_k \gg 1$. As for biological necessity to ensure proper replication, median RF stalling distance N_s needs to be much larger than N_k, the distance between adjacent pair of ROs k and $k+1$. Indeed the available experimental suggestions for N_s is in the scale

of genome length in yeasts and thus few hundred ROs spread out over the genome intuitively will produce much lesser inter-RO distances than N_s. At the same time N_k cannot be smaller than the molecular footprint of the protein complexes that licenses the ROs e.g. MCM2-7 double hexamer has a footprint of about 60 base pairs. Thus N_k is necessarily bigger than 1.

Probability of DFS between two adjacent ROs

We recall the general formula from the previous section for the probability of DFS in the region of DNA between two adjacent ROs, namely Eq. (6)

$$P(\text{DFS in } D) \;=\; 1-(1+Nq)(1-q)^N$$

Since, q is very small, $Nq \;\ll 1$. Using binomial expansion, we have

$$(1-q)^N \;\approx\; 1-Nq+\frac{1}{2}\,q^2N(N-1)+O[(Nq)^3].$$

Thus,

$$P(\text{DFS in } D) \;=\; 1-(1+Nq)\left(1-Nq+\frac{1}{2}\,q^2N(N-1)\right)$$

$$=\; 1-\left(1-N^2q^2+\frac{1}{2}\,q^2N(N-1)\right)$$

$$=\; N^2q^2-\frac{1}{2}\,q^2N(N-1)$$

This gives us the following simple exact result (to order $(Nq)^2$):

$$P(\text{DFS in } D)=\frac{q^2}{2}N(N+1) \qquad (A1).$$

Since the typical distance between licensed ROs $N_l \gg 1$, we can simplify this exact result to

$$P(\text{DFS in } D) = \frac{(qN)^2}{2} \qquad (A2).$$

By the definition of N_s (the 'median stalling distance'), we have

$$P(\text{fork starting at position 0 and stalling at any } n \geq N_s) = \frac{1}{2} \qquad (A3).$$

Let us denote this long-winded probability by $P(\text{median stall})$. Now, according to Eq. (2), we have

$$P(\text{stall at any } n < N_s) = \sum_{n=0}^{N_s-1} (1-q)^n q \,,$$

$$= 1 - (1-q)^{N_s}$$

So,

$$P(\text{no stall at any } n < N_s \text{ or stall at any } n \geq N_s) = (1-q)^{N_s}$$

Which means,

$$P(\text{median stall}) = (1-q)^{N_s}$$

According to Eq. (*A*3) we now have an exact relationship between q and N_s:

$$\frac{1}{2} = (1-q)^{N_s}$$

Taking natural logarithms, we have:

$$-\log 2 = N_s \log(1-q)$$

Since $q \ll 1$, $\log(1-q) \approx -q$, and thus we derive the following expression

$$q = \frac{\log(2)}{N_s} \qquad (A4).$$

We can use Eq. (*A*4) to write Eq. (*A*2) purely in terms of N_s and we get

$$P(\text{DFS in } D) = \frac{(\log 2)^2}{2} \left(\frac{N}{N_s}\right)^2$$

Defining the constant $\alpha = \frac{(\log 2)^2}{2} \approx 0.240\ldots$ and expressing '*D*' in terms of its nucleotide content in base pairs, *N,* we have

$$p_{\mathrm{DFS}}(N) = \alpha \left(\frac{N}{N_s}\right)^2 \qquad (A5).$$

Spatial variation in RO distribution

We define the inter-RO distance between adjacent ROs labelled by $(i+1)$ and i by N_i. Now, associated with this pair of ROs is the probability of a DFS, $p_{\mathrm{DFS}}(N_i)$ and we denote this by p_i, just for convenience. So, we have

$$p_{\mathrm{DFS}}(N_i) = p_i = \alpha \left(\frac{N_i}{N_s}\right)^2 \qquad (A6).$$

Now, we denote the probability of no DFS genome-wide by $P(0\ \mathrm{DFS})$, which is simply given by the following product of independent probabilities for no DFS in every possible region of separation between adjacent ROs:

$$P(0\ \mathrm{DFS}) = (1-p_1)(1-p_2)\ldots(1-p_i)\ldots$$

Or,

$$P(0\ \mathrm{DFS}) = \prod_i (1-p_i) \qquad (A7).$$

Using the fact that a product of factors can be rewritten as the exponential of a sum of logarithms of these factors, we can rewrite the above equation in the following form

$$P(0\ \mathrm{DFS}) = \exp \log \prod_i (1-p_i) = \exp \sum_i \log (1-p_i)$$

Now, since we have assumed that $1 \ll N_i \ll N_s$ for all i, the value of $p_{\mathrm{DFS}}(N_i)$ or p_i which is $\alpha \left(\frac{N_i}{N_s}\right)^2$, implies that $p_i \ll 1$. Thus $\log(1-p_i) \approx -p_i$ and above equation takes the following simpler form

$$P(0\ \mathrm{DFS}) = \exp\left(-\sum_i p_i\right) \qquad (A8).$$

We define an average of the independent quantities $\{p_{\mathrm{DFS}}(N_i)\}$ or $\{p_i\}$, and their overall number. We denote the average by $\langle p_i \rangle$. The overall number is the size of the

genome divided by the average inter-RO distance (denoted by N_l), that is (approximately) N_g/N_l. Then the law of large numbers provides us with the relation:

$$\sum_i p_{\mathrm{DFS}}(N_i) = \frac{N_g}{N_l}\,\langle p_i\rangle \qquad (A9).$$

But, as we know $p_i = \alpha\left(\frac{N_i}{N_s}\right)^2$, we can directly relate $\langle p_i\rangle$ to the second moment of inter-RO distance N_i i.e.

$$\langle p_i\rangle = \alpha\,\langle\left(\frac{N_i}{N_s}\right)^2\rangle = \alpha\,\frac{\langle {N_i}^2\rangle}{{N_s}^2} \qquad (A10).$$

Now, using Eq. (A10), we rewrite Eq. (A9) as below

$$\sum_i p_{\mathrm{DFS}}(N_i) = \alpha\,\frac{N_g}{N_l}\,\frac{\langle {N_i}^2\rangle}{{N_s}^2}$$

So, it is clear to write Eq. (A8) as:

$$P(0\ \mathrm{DFS}) = \exp\left(-\alpha\,\frac{N_g\langle {N_i}^2\rangle}{{N_s}^2 N_l}\right) \qquad (A11).$$

The second moment of a distribution is equal to the square of the mean plus the variance. So, denoting the variance in the inter-RO separation by $var(N_k)$, we have

$$\langle {N_i}^2\rangle = \langle N_i\rangle^2 + var(N_i)$$

By definition, $\langle N_i\rangle = N_l$ and so we write Eq. (A11) more explicitly in terms of variance:

$$P(0\ \mathrm{DFS}) = \exp\left(-\alpha\,\frac{N_g N_l}{{N_s}^2}\left(1 + \frac{\mathrm{var}\,(N_i)}{{N_l}^2}\right)\right)$$

By replacing the variance with the standard deviation of the inter-RO distances $sd(N_i)$, we have

$$P(0\ \mathrm{DFS}) = \exp\left(-\alpha\,\frac{N_g N_l}{{N_s}^2}\left(1 + \left[\frac{sd(N_i)}{N_l}\right]^2\right)\right)$$

We denote the ratio of standard deviation to mean, $sd(N_i)/N_l$, as R and thus we have

$$P(0\ \text{DFS}) = \exp\left(-\alpha\frac{N_g N_l}{{N_s}^2}(1+R^2)\right)$$

Now,

$$P(\text{DFS}) = 1 - P(0\ \text{DFS}),$$

and so,

$$P(\text{DFS}) = 1 - \exp\left(-\alpha\frac{N_g N_l}{{N_s}^2}(1+R^2)\right) \qquad (A12).$$

In the event that this probability is very small, meaning $P(\text{DFS}) \ll 1$, in which case the argument of the exponential must be small, and we have our main result

$$P(\text{DFS}) \approx \alpha\frac{N_l N_g}{{N_s}^2}(1+R^2) \qquad (A13).$$

Error from the largest origin separation

We denote the largest gap between adjacent ROs in the given dataset by $N_{\max}$. Now, from Eq. (*A6*), we can directly write the probability of DFS in the specific inter-RO separation denoted by $N_{\max}$, as following

$$P(\text{DFS in } N_{\max}) = \alpha\left(\frac{N_{\max}}{N_s}\right)^2 \qquad (A14)$$

Errors at chromosome ends

We have already mentioned that very end of the chromosome represents a special case. This very end from the end-proximal RO to the telomeric end is replicated by a single RF and a single stall of this terminal RF is sufficiently enough to cause a similar catastrophe like DFS, essentially to cause a portion of DNA to remain unreplicated. According to Eq. (2), we write

$$P(\text{fork starting at position 0 and stalling at any } n < N) = (1-q)^n q$$

Or probability of a 'telomeric single fork stall' (TFS) at any *n* within a specified region is simply

$$p_{\mathrm{TFS}}(n) = (1-q)^n q$$

Thus, for a chromosome end, which has a length of n_e in base pairs, the SFS probability for a RF can be given as

$$P(\text{TFS at end or } n < n_e) = \sum_{n=0}^{n_e-1} (1-q)^n q$$

Using Eq. (5), we get

$$P(\text{TFS at end}) = 1 - (1-q)^{n_e}$$

So, we can write

$$P(\text{no TFS at a chromosome end}) = (1-q)^{n_e} \qquad (A15).$$

Let us consider the total length of all chromosome ends is N_e. Now, the probability of no SFS error at any of the chromosome ends is

$$P(\text{no TFS at chromosome ends}) = (1-q)^{N_e} \qquad (A16).$$

We can write this product in the form of exponential, using natural logarithms

$$P(\text{no TFS at chromosome ends}) = \exp\left[N_e \log(1-q)\right]$$

Since q is very small, $\log(1-q) \approx -q$ and we write

$$P(\text{no TFS at chromosome ends}) = \exp\left(-qN_e\right) \qquad (A17).$$

Now, we use Eq. ($A4$) to rewrite Eq. ($A17$) in terms of N_s

$$P(\text{no TFS at chromosome ends}) = \exp\left(-\frac{(\log 2)\, N_e}{N_s}\right)$$

So, it is now straightforward to write,

$$P(\text{TFS at chromosome ends}) = 1 - \exp\left(-\frac{(\log 2)\, N_e}{N_s}\right)$$

Since, $P(\text{TFS at chromosome ends}) \ll 1$, the argument of the exponential must be small, and we have

$$P(\text{TFS at chromosome ends}) \approx \frac{(\log 2)\, N_e}{N_s} \qquad (A18).$$

Estimating median stalling distance

Replication failures at the extreme end of chromosome arising from TFS and at the bulk chromosome from DFS should be similar in order to balance the error during replication. Hence, we write

$$P(\text{TFS at } N_e) \approx P(\text{DFS})$$

Which means, comparing Eq. (*A*13) with Eq. (*A*18) we have

$$\frac{(\log 2)N_e}{N_s} \approx \alpha \frac{N_l N_g}{{N_s}^2}(1+R^2)$$

Simply by considering $\alpha = \frac{(\log 2)^2}{2}$, and slightly rearranging above expression we have the following expression for N_s,

$$N_s \approx \left(\frac{\log 2}{2}\right)\frac{N_l N_g}{N_e}(1+R^2) \qquad (A19).$$

Which provides a straightforward prediction for N_s (or alternatively for q according to Eq. (*A*4)) from known genomic quantities.

2.4 Model B

In **Model A** we assumed the inter-RO distances are much smaller than N_s, which is biologically found in smaller genomes like in yeasts, this allowed us to make approximations used to derive the formulas in there. For larger genomes in higher eukaryotes, this assumption is lifted due to the possibility of large variability in inter-RO distances over the very big genomes with hundreds of thousands of ROs. Hence, in **Model B** we leave the approximations used before and provide the more general derivations below.

Derivation of the 'central equation' for probability of DFS

The general formula for the probability of a DFS in the region of DNA between two adjacent ROs, given by Eq. (6) is rewritten as

$$P(\text{DFS in } D) \;=\; 1-(1+Nq)(1-q)^N \qquad (B1).$$

Expressing the product as the exponential of the sum of the logarithms gives

$$(1-q)^N = \exp(N\log(1-q)) \qquad (B2).$$

Since q is an extremely small number, $\log(1-q) \approx -q$, and hence

$$(1-q)^N = \exp(-Nq) \qquad (B3).$$

Combining Eq. ($B1$) with Eq. ($B3$), we obtain

$$P(\text{DFS in } D) \;=\; 1-(1\;+\;Nq)\cdot\exp(-Nq) \qquad (B4).$$

Let us define the distance between the adjacent $(i+1)^{\text{th}}$ and i^{th} ROs as N_i. The probability of DFS between this pair of ROs will be denoted as P_i. Thus,

$$P_i \;=\; 1-(1\;+\;N_iq)\exp(-N_iq) \qquad (B5).$$

The genome-wide probability of no double stall or 0 DFS, which will be denoted as $P(0\text{ DFS})$, is given by the product of probability of no double stalls in each inter-RO gap, i.e.

$$P(0\text{ DFS}) \;=\; \prod_i (1-\;P_i) \qquad (B6).$$

Combining Eq. ($B5$) and Eq. ($B6$), we have

$$P(0\text{ DFS}) \;=\; \left\{\prod_i (1\;+\;N_iq)\right\}\left\{\prod_i (\exp(-N_iq))\right\} \qquad (B7).$$

Let N_g be the genome length, then

$$\sum_i N_i = N_g$$

Thus

$$\prod_i (\exp(-N_i q)) = \exp\left(-q \sum_i N_i\right) = \exp(-qN_g) \qquad (B8).$$

Similarly,

$$\prod_i (1 + N_i q) = \prod_i \exp(\log(1 + N_i\, q)),$$

$$\prod_i (1 + N_i q) = \exp\left(\sum_i \log(1 + N_i\, q)\right) \qquad (B9).$$

Hence, we obtain the following by combining Eqs. $(B7)$, $(B8)$ and $(B9)$

$$P(0\ \text{DFS}) = \exp(-q\, N_g)\ \exp\left(\sum_i \log(1 + N_i\, q)\right)$$

Or

$$P(0\ \text{DFS}) = \exp\left(-q\, N_g + \sum_i \log(1 + N_i\, q)\right)$$

In **Model A**, we have shown that per-nucleotide stalling rate for a RF, $q = \log(2)/N_s$. So

$$P(0\ \text{DFS}) = \exp\left(-\frac{\log(2)\, N_g}{N_s} + \sum_i \log\left(1 + \frac{\log(2)\, N_i}{N_s}\right)\right) \qquad (B10).$$

Probability of specific number of DFSs

Probability of an arbitrary number of DFSs can be calculated by extending the previous approach. The probability of exactly 1 DFS, which will be called $P(1\ \text{DFS})$, can be calculated directly as

$$P(1\ \text{DFS}) = \sum_i P_i \prod_{i_1 \neq i} (1 - P_{i_1}) \qquad (B11).$$

Combining Eq. $(B11)$ with Eq. $(B6)$ we obtain

$$P(1\ \text{DFS}) = P(0\ \text{DFS}) \sum_i \frac{P_i}{(1 - P_i)}$$

Therefore

$$\frac{P(1\ \text{DFS})}{P(0\ \text{DFS})} = \sum_i \frac{P_i}{(1 - P_i)}$$

To simplify the next steps, we introduce the following definitions

$$\mathrm{S}_1 = \sum_i \frac{P_i}{1 - P_i}$$

$$\mathrm{S}_2 = \sum_i \left(\frac{P_i}{1 - P_i}\right)^2$$

$$\vdots$$

$$\mathrm{S}_m = \sum_i \left(\frac{P_i}{1 - P_i}\right)^m$$

Additionally, let $P(m\ \text{DFS})$ be the probability of m DFS, the following conventions will be used

$$R_1 = \frac{P(1\ \text{DFS})}{P(0\ \text{DFS})}$$

$$R_2 = \frac{P(2\ \text{DFS})}{P(0\ \text{DFS})}$$

$$\vdots$$

$$R_m = \frac{P(m\ \text{DFS})}{P(0\ \text{DFS})} \qquad (B12).$$

Hence,

$$R_1 = \sum_i \frac{P_i}{(1 - P_i)} = S_1 \qquad (B13).$$

And

$$R_2 = \frac{1}{2!} \sum_{i_1} \sum_{i_2 \neq i_1} \frac{P_{i_1}}{(1 - P_{i_1})} \frac{P_{i_2}}{(1 - P_{i_2})}$$

which can be rewritten as

$$R_2 = \frac{1}{2!} \left[\left(\sum_i \frac{P_i}{1 - P_i} \right)^2 - \sum_i \left(\frac{P_i}{1 - P_i} \right)^2 \right] = \frac{1}{2!} \left(S_1^2 - S_2 \right)$$

Similarly

$$R_3 = \frac{1}{3!}\sum_{i_1}\sum_{i_2\neq i_1}\sum_{i_3\neq i_1,i_2}\frac{P_{i_1}}{(1-P_{i_1})}\frac{P_{i_2}}{(1-P_{i_2})}\frac{P_{i_3}}{(1-P_{i_3})}$$

$$= \frac{1}{3!}\,(S_1^3 - 3\,S_1\,S_2 + 2\,S_3)$$

Iterating the same approach it is possible to show that

$$R_4 = \frac{1}{4!}\,(S_1^4 - 6\,S_1^2\,S_2 + 8\,S_1\,S_3 + 3\,S_2^2 - 6\,S_4)$$

$$R_5 = \frac{1}{5!}\,(S_1^5 - 10\,S_1^3\,S_2 + 15\,S_1\,S_2^2 + 20\,S_1^2\,S_3 - 20\,S_2\,S_3 - 30\,S_1\,S_4 + 24\,S_5)$$

$$R_6 = \frac{1}{6!}\,(S_1^6 - 15\,S_1^4 S_2 + 45\,S_1^2 S_2^2 - 15\,S_2^3 + 40\,S_1^3 S_3 - 120\,S_1 S_2 S_3$$

$$+ 40\,S_3^2 - 90\,S_1^2 S_4 + 90\,S_2 S_4 + 144\,S_1 S_5 - 120\,S_6)$$

Finally, combining R_1, R_2, R_3, R_4, R_5 and R_6 with Eq. $(B12)$, we can obtain the probability of one to six double fork stalls as follows

$$P(m\ \text{DFS}) = P(0\ \text{DFS}) \cdot R_m$$

Distribution of DFSs follows Poisson

This section is discussed in more details in Chapter 4, section 4.3.5. Here, as shown in the following Table, direct calculations on the human cell line IMR90 show that only the leading power is playing a significant role for the value of R_k.

Value used for 1 DFS	S1	1.68
Values used for 2 DFS	$S1^2$	2.81
	$S2/S1^2$	$8.84 \cdot 10^{-4}$
Values used for 3 DFS	$S1^3$	4.71
	$S1*S2/S1^3$	$8.84 \cdot 10^{-4}$
	$S3/S1^3$	$9.25 \cdot 10^{-6}$
Values used for 4 DFS	$S1^4$	7.89
	$S1^2 * S2/S1^4$	$8.84 \cdot 10^{-4}$
	$S1*S3/ S1^4$	$9.25 \cdot 10^{-6}$
	$S2^2/ S1^4$	$7.81 \cdot 10^{-7}$
	$S4/ S1^4$	$1.40 \cdot 10^{-7}$

Therefore, we have:

$$R_2 \approx \frac{1}{2!}\,(S_1^2)$$

$$\vdots$$

$$R_k \approx \frac{1}{k!}\,\left(S_1^k\right)$$

and hence

$$P(k\,\mathrm{DFS}) \;=\; P(0\,\mathrm{DFS})\;\cdot R_k \approx\; P(0\,\mathrm{DFS})\frac{1}{k!}\,\left(S_1^k\right)$$

which indicates a Poisson distribution.

The probability density function of a Poisson distribution is

$$P(\mathrm{n}) = \exp(-\lambda)\frac{\lambda^n}{n!}$$

So, for our distribution to follow a Poisson we have to show that

$$P(0\,\mathrm{DFS}) = \exp(-\lambda) \qquad (B14)$$

which implies

$$\mathrm{S}_1 = \lambda.$$

From ($B6$) and ($B14$) we have

$$\lambda = -\log\left(\prod_i (1-P_i)\right) = -\sum_i \log(1-P_i) = \sum_i \log\left(\frac{1}{1-P_i}\right)$$

$$= \sum_i \log\left(1+\frac{P_i}{1-P_i}\right)$$

The value $P_i/(1-P_i)$ is very small and we use a Taylor expansion to obtain

$$\lambda = \sum_i \left(\frac{P_i}{1-P_i} + O(P_i^2)\right) \approx \sum_i \left(\frac{P_i}{1-P_i}\right) = S_1$$

Since the P_i is very small, this approximation is generally very good.

Frequency of inter-RO gaps of a particular size

The inter-RO gaps in the human cell lines vary widely ranging from very small to very large size. In order to check the relative contribution of the different size ranges among the gaps to the overall genomic error rate, we want to calculate the frequency of gaps in particular size ranges. This will be used to quantify their relative contribution of genomewide error in chapter 4, section 4.3.4.

In Eq. ($B1$), we have shown that the probability of a DFS inside the region of DNA between a pair of adjacent ROs separated by N nucleotides is,

$$P(N) = 1 - (1 + Nq)(1 - q)^N \qquad (B15).$$

Now, we calculate the probability of DFS in a cohort of M gaps whose size is "in the vicinity of" N. The probability of no error occurring from this cohort would be the following product,

$$\prod_i (1 - P(N_i)).$$

where the product is restricted to those gaps within the cohort. This probability will be very close to one, and we denote it by θ. Substituing ($B15$) into this expression, and recognizing that all of the N_k are close to N, enables us to rewrite the probability of no error from the cohort as:

$$(1 + Nq)^M (1 - q)^{MN} = \theta$$

Now, taking the natural logarithm,

$$M \cdot \log(1 + Nq) + MN \cdot \log(1 - q) = \log(\theta) \qquad (B16).$$

Since q << 1, $\log(1 - q) \approx -q$, and thus we write

$$M \cdot \log(1 + Nq) - MNq = \log(\theta),$$

Hence,

$$M = \frac{\log(\theta)}{[\log(1+Nq) - Nq]} \qquad (B17).$$

For $Nq \ll 1$, Taylor expansion of $\log(1+Nq) \approx Nq - \frac{1}{2}N^2q^2 + O[(Nq)^3] \approx Nq - \frac{1}{2}N^2q^2$. So, we have

$$M = \frac{\log(\theta)}{\left[Nq - \frac{1}{2}N^2q^2 - Nq\right]}$$

$$M = \frac{\log\left(\frac{1}{\theta}\right)}{\frac{1}{2}N^2q^2}$$

or,

$$M = \frac{2\log\left(\frac{1}{\theta}\right)}{q^2}\frac{1}{N^2}.$$

Since, $\frac{2\log\left(\frac{1}{\theta}\right)}{q^2}$ constitutes a mathematical constant, we have

$$M \sim \frac{1}{N^2} \qquad (B18).$$

2.5 Model C

In eukaryotic genomes, the minimum inter-RO length is bounded by the structure of nucleosomes and replication machineries. The distance between two adjacent ROs cannot be less than the sum of the lengths of histone core of the nucleosome and the effective footprint of RO licensing factors. Simply because ROs cannot be licensed in the wrapped histone cores in nucleosomes rather can only be licensed in the linker segments. Hence, in '**Model C**' we consider each nucleosome linker as a potential site for RO licensing. We rewrite the general formula for the probability of genome wide zero DFSs or no error as given in Model A as Eq. (A7) and in Model B as Eq. (B6).

$$P(0\ \text{DFS}) = P(\text{no error}) = \prod_i (1 - p_i)$$

$$P(\text{no error}) = \exp\left[\sum_i \log(1 - p_i)\right]$$

since, $p_i \ll 1$, $\log(1 - p_i) = -p_i$, so

$$P(\text{no error}) = \exp\left[-\sum_i p_i\right]$$

hence,

$$P(\text{error}) = 1 - \exp\left[-\sum_k p_k\right] \qquad (C1).$$

Let us define, ρ as probability of licensed RO in nucleosome. Thus, probability of having a gap of *n* nucleosomes is G_n where

$$G_n = \rho(1 - \rho)^{n-1}$$

and, the probability of error in G_n is given by

$$p_n = 1 - (1 + nN_{nuc}q)(1 - q)^{nN_{nuc}}$$

the total number of ROs in a given genome carrying '*M*' nucleosomes is given by

$$Number\ of\ ROs = M\rho$$

So, the overall number of gaps of size *n* nucleosomes in the genome is

$$G = M\rho\,.\rho(1 - \rho)^{n-1} = M\rho^2(1 - \rho)^{n-1}$$

For, genomewide *i* gaps of size *n* nucleosomes,

$$\sum_i p_i = \sum_n G\,p_n = \sum_{n=1}^{\infty} G\,p_n \qquad (C2).$$

Now,

$$\sum_{n=1}^{\infty} G\,p_n = \sum_{n=1}^{\infty} M\rho^2(1 - \rho)^{n-1}\,[1 - (1 + nN_{nuc}q)(1 - q)^{nN_{nuc}}]$$

$$= M\rho^2 \sum_{n=1}^{\infty}(1 - \rho)^{n-1}\ \ [1 - (1 - q)^{nN_{nuc}} - nN_{nuc}q(1 - q)^{nN_{nuc}}]$$

$$= M\rho^2\left[\sum_{n=1}^{\infty}(1-\rho)^{n-1} - \sum_{n=1}^{\infty}(1-\rho)^{n-1}\{(1-q)^{N_{nuc}}\}^n - N_{nuc}q\sum_{n=1}^{\infty}(1-\rho)^{n-1}\{(1-q)^{N_{nuc}}\}^n\right]$$

We simplify as,

$$\sum_{n=1}^{\infty}(1-\rho)^{n-1} = \frac{1}{1-(1-\rho)} = \frac{1}{\rho};$$

$$\sum_{n=1}^{\infty}(1-\rho)^{n-1}z^n = \frac{z}{1-z(1-\rho)};$$

$$\sum_{n=1}^{\infty}(1-\rho)^{n-1}n\,z^n = \frac{z}{[1-z(1-\rho)]^2}.$$

Using these simplifications we have,

$$\sum_{n=1}^{\infty}G\,p_n = M\rho^2\left[\frac{1}{\rho} - \frac{(1-q)^{N_{nuc}}}{1-(1-\rho)(1-q)^{N_{nuc}}} - N_{nuc}q\frac{(1-q)^{N_{nuc}}}{[1-(1-\rho)(1-q)^{N_{nuc}}]^2}\right]$$

$$= M\rho^2\left[\frac{[1-(1-\rho)(1-q)^{N_{nuc}}]^2 - \rho[1-(1-\rho)(1-q)^{N_{nuc}}](1-q)^{N_{nuc}} - \rho N_{nuc}q(1-q)^{N_{nuc}}}{\rho\,[1-(1-\rho)(1-q)^{N_{nuc}}]^2}\right],$$

or,

$$= M\rho\left[\frac{[1-(1-\rho)(1-q)^{N_{nuc}}]^2 - \rho[1-(1-\rho)(1-q)^{N_{nuc}}](1-q)^{N_{nuc}} - \rho N_{nuc}q(1-q)^{N_{nuc}}}{[1-(1-\rho)(1-q)^{N_{nuc}}]^2}\right] \quad (C3).$$

Now, we make the following binomial expressions,

- $(1-q)^{N_{nuc}} \approx 1 - qN_{nuc} + \frac{1}{2}(qN_{nuc})^2$
- $1-(1-\rho)(1-q)^{N_{nuc}} \approx 1-(1-\rho)\left(1 - qN_{nuc} + \frac{1}{2}(qN_{nuc})^2\right)$

  $$= \rho + (1-\rho)qN_{nuc} - \frac{1}{2}(1-\rho)(qN_{nuc})^2$$

- $[1-(1-\rho)(1-q)^{N_{nuc}}]^2 = \left[\rho + (1-\rho)qN_{nuc} - \frac{1}{2}(1-\rho)(qN_{nuc})^2\right]^2$

Putting these values inside Eq. (*C*3) and after doing the algebra,

$$\sum_{n=1}^{\infty} G\, p_n = M\rho \left[\frac{\left(1 - \frac{\rho}{2}\right)(qN_{nuc})^2}{\left[\rho + (1 - \rho)qN_{nuc} - \frac{1}{2}(1 - \rho)(qN_{nuc})^2\right]^2} \right]$$

Since, q is very small, $qN_{nuc} \ll 1$ and hence the leading term in the denominator is ρ^2. Thus we have,

$$\sum_{n=1}^{\infty} G\, p_n = M\rho \left[\frac{\left(1 - \frac{\rho}{2}\right)(qN_{nuc})^2}{\rho^2} \right] = \frac{M}{\rho} \frac{(2 - \rho)q^2 {N_{nuc}}^2}{2}$$
$$= \frac{M\, {N_{nuc}}^2}{\rho} \frac{(2 - \rho)}{2} \left(\frac{\log(2)}{N_s}\right)^2 \qquad (C4).$$

We know, by definition $M = \frac{N_g}{N_{nuc}}$. Replacing *M* in Eq. (*C*4), we have

$$\sum_{n=1}^{\infty} G\, p_n = \frac{N_g\, N_{nuc}}{\rho} \frac{(2 - \rho)}{{N_s}^2} \frac{(\log(2))^2}{2}$$

We know from Eq. (*A*4), per nucleotide fork stall rate

$$q = \frac{\log(2)}{N_s}.$$

Thus,

$$\sum_{n=1}^{\infty} G\, p_n = \frac{q^2 N_{nuc}}{2} \frac{N_g(2 - \rho)}{\rho}$$

hence,

$$\sum_i p_i = \frac{q^2 N_{nuc}}{2} \frac{N_g(2 - \rho)}{\rho}$$

We rewrite Eq. (*C*1) as

$$P(\text{error}) = 1 - \exp\left[-\frac{q^2 N_{nuc}}{2} \frac{N_g(2 - \rho)}{\rho} \right] \qquad (C5).$$

As q and N_{nuc} are biologically conserved factors across eukaryotes, we define them as the constant U and thus

$$\frac{q^2 N_{nuc}}{2} = U$$

So,

$$P(\text{error}) = 1 - \exp\left[-U\,\frac{N_g(2-\rho)}{\rho}\right] \qquad (C6).$$

In order to calculate U back from the experimentally observed replication error rates; we rearrange Eq. (*C*6),

$$\exp\left[-U\,\frac{N_g(2-\rho)}{\rho}\right] = 1 - P(\text{error});$$

$$-U\,\frac{N_g(2-\rho)}{\rho} = \log\{1 - P(\text{error})\} \qquad (C7).$$

Or,

$$U\,\frac{N_g(2-\rho)}{\rho} = \log\left\{\frac{1}{1-P(\text{error})}\right\};$$

For convenience, here we define $U = U_{calculated}$ and we have

$$U_{calculated} = -\frac{\rho}{(2-\rho)}\,\frac{\log\big(1-P(\text{error})\big)}{N_g} \qquad (C8).$$

For non-embryo like systems, ρ is less than 1 and $P(\text{error})$ could be high depending on the genome size of the organism; so Eq. (*C*8) is applicable to such systems.

Let us rewrite the Eq. (*C*7)

$$-U\,\frac{N_g(2-\rho)}{\rho} = \log\{1 - P(\text{error})\}$$

When, $\rho \to 1;\ P(\text{error}) \ll 1$, hence

$$\log\{1 - P(\text{error})\} = -P(\text{error})$$

Thus Eq. (*C*6), in this context takes the following simpler form

$$-U\,N_g = -P(\text{error})$$

or,

$$U = \frac{P(\text{error})}{N_g} \qquad (C9).$$

In early embryo $\rho \to 1$, hence Eq. (*C*9) is suitable for such systems. As in early embryos, replication is the only genomic activity hence embryonic mortality at very

early stages of development could be used to estimate $P(\text{error})$ in embryos. Due to the fact that embryo mortality could be resulted from additional factors other than replication, so the error estimated from the observed mortality $P_{observed}$ could be greater than $P(\text{error})$ which is the measure of replication error only. So, for

$$P_{observed} \geq P(\text{error})$$

we have,

$$\frac{P_{observed}}{N_g} \geq \frac{P(\text{error})}{N_g}.$$

and thus,

$$\frac{P_{observed}}{N_g} \geq U.$$

Convergence of Model C to Model A

With ρ as the probability of licensed RO in nucleosome and M as the number of nucleosomes in a given genome, the total number of RO on the genome is given by $M\rho$. We express the mean inter-RO distance N_l as following:

$$N_l = \frac{N_g}{M\rho} = \frac{N_g}{\frac{N_g}{N_{nuc}}\,\rho} = \frac{N_{nuc}}{\rho}$$

For convenience, we rewrite the Eq. ($C5$) here

$$P(\text{error}) = 1 - \exp\left[-\frac{q^2 N_{nuc}}{2}\,\frac{N_g(2-\rho)}{\rho}\right] \qquad (7).$$

Replacing $\frac{N_{nuc}}{\rho}$ with N_l we have

$$P(\text{error}) = 1 - \exp\left[-\frac{q^2 N_l\, N_g(2-\rho)}{2}\right] \qquad (8).$$

In Model A, we have shown that $q = \frac{\log(2)}{N_s}$ and defined the constant $\alpha = \frac{(\log 2)^2}{2}$.

Replacing q and taking the constant α in Eq. (8) gives

$$P(\text{error}) = 1 - \exp\left[-\alpha\,\frac{N_l\, N_g}{{N_s}^2}\,(2-\rho)\right] \qquad (9).$$

The overall number of gaps of size *n* nucleosomes in the genome is

$$G = M\rho\,.\rho(1-\rho)^{n-1} = M\rho^2(1-\rho)^{n-1}$$

Now,

$$\sum_{n=1}^{\infty} G = M\rho$$

$$\sum_{n=1}^{\infty} Gn = M\rho^2 \sum_{n=1}^{\infty} n(1-\rho)^{n-1}$$

$$= M\rho^2 \sum_{n=1}^{\infty} n(1-\rho)^{n-1}$$

$$= M\rho^2(-\partial\rho) \sum_{n=1}^{\infty} (1-\rho)^n = M\rho^2(-\partial\rho)\frac{(1-\rho)}{\rho} = M$$

and

$$\sum_{n=1}^{\infty} Gn^2 = M\rho^2 \sum_{n=1}^{\infty} n^2(1-\rho)^{n-1}$$

$$= M\rho^2(-\partial\rho) \sum_{n=1}^{\infty} n(1-\rho)^n$$

$$= M\rho^2(-\partial\rho)(1-\rho)\,(-\partial\rho) \sum_{n=1}^{\infty} (1-\rho)^n$$

$$= M\rho^2(-\partial\rho)(1-\rho)\,(-\partial\rho)\frac{(1-\rho)}{\rho} = M\rho^2(-\partial\rho)(1-\rho)\frac{1}{\rho^2}$$

$$= M\rho^2(-\partial\rho)\left(\frac{1}{\rho^2} - \frac{1}{\rho}\right) = M\rho^2\left(\frac{2}{\rho^3} - \frac{1}{\rho^2}\right) = M\left(\frac{2}{\rho} - 1\right)$$

Now,

$$\sum_{n=1}^{\infty} Gn \div \sum_{n=1}^{\infty} G = \frac{M}{M\rho} = \frac{1}{\rho} = \langle n \rangle$$

$$\sum_{n=1}^{\infty} Gn^2 \div \sum_{n=1}^{\infty} G = \frac{M\left(\frac{2}{\rho} - 1\right)}{M\rho} = \frac{\left(\frac{2}{\rho} - 1\right)}{\rho} = \langle n^2 \rangle$$

We know, coefficient of variation $R = \frac{sd}{mean}$ or $R^2 = \frac{variance}{mean^2}$ … and we know variance of a distribution is second moment minus the square of the mean, thus

$$R^2 = \frac{\langle n^2 \rangle - \langle n \rangle^2}{\langle n \rangle^2} = \frac{\langle n^2 \rangle}{\langle n \rangle^2} - 1 = \frac{\frac{\left(\frac{2}{\rho} - 1\right)}{\rho}}{\left(\frac{1}{\rho}\right)^2} - 1 = 1 - \rho$$

Hence,

$$2 - \rho = 1 + R^2$$

replacing $(2 - \rho)$ in Eq. (9) we have

$$P(\text{error}) = 1 - \exp\left[-\alpha \frac{N_l N_g}{{N_s}^2} (1 + R^2)\right] \qquad (10).$$

which is Eq. ($A12$), the formula for probability of DFS error genomewide in Model A.

Chapter 3

Regular RO distribution minimizes the replication error in yeasts

3.1 Brief Introduction

Eukaryotic genomes are replicated during S phase of the cell cycle by activating hundreds or thousands of replication origins (ROs), which are licensed through recruitment of minichromosome maintenance proteins (MCM2-7) before replication initiation. It is fundamentally important for the cells to avoid double replication of any DNA segment as well as to ensure no portion of DNA remains unreplicated. This is achieved, firstly by restricting the licensing of ROs before replication begins and secondly by passive replication of inactive ROs i.e. inactive MCMs are removed from DNA when visited by an active replication fork (RF) from any other activated RO (Arias and Walter, 2007; Blow and Dutta, 2005). Hence, the number of ROs licensed must be sufficient enough and it is a crucial matter in confirming complete replication of the eukaryotic genome. More importantly the irreversible stalling of the RFs i.e. DFS (double fork stall) and TFS (telomeric fork stall) discussed in chapter 1 (Figure 2), makes the number of licensed ROs a major constraint for complete replication (Cobb et al., 2005; De Piccoli et al., 2012). Experimentally it is being confirmed that indeed eukaryotes license almost three times excess to the number of ROs actively needed to complete the whole genome duplication and alongside the number, the distribution of licensed ROs is also very important in order to finish complete replication (Blow et al., 2011; Blow and Ge, 2009; Ge et al., 2007; Woodward et al., 2006). Even though the RO positions in eukaryotes are long studied and at least in budding yeast these positions are well documented for quite some time now yet not

much is known about the system level constraints and biological features that are responsible in establishing the number of licensed ROs and their distribution (Borowiec and Schildkraut, 2011; Gilbert, 2012; Nieduszynski et al., 2006). Organisms like yeasts with haploid genome length of ~10 Mbp license a few hundred ROs spread-out over the whole genome. In budding yeast, ROs are licensed in a sequence specific manner while in fission yeast ROs are not strictly sequence specific (Barberis et al., 2010; Patel et al., 2006). This portrays a basic importance of the number and distribution of licensed ROs in these organisms irrespective of the divergence in biological detail governing RO licensing. Due to the fact that few hundred ROs are licensed over the genome of ~10 Mbp length, intuitively the possible inter-RO distances cannot be very large in the order of Mbp or such. Hence, **Model A** is appropriate for this genome length where we assume all the inter-RO distances are much smaller than 'median stalling distance' N_s which is experimentally estimated to be similar to the genome length in yeasts i.e. ~10 Mbp (Maya-Mendoza et al., 2007). In this chapter, we are going to apply **Model A** in five different yeast species in order to decipher the constraints governing the distribution of licensed ROs in their genomes. Appropriate formulas will be cited from chapter 2 accordingly and main results, which are published in the journal Nucleic Acid Research (**NAR**) (Newman et al., 2013), will be discussed below.

3.2 Data for RO distribution in yeasts

Saccharomyces cerevisiae RO positions were selected based on the data at OriDB using the following criteria:

a) ROs that have been experimentally determined by an Autonomously Replicating Sequence (ARS) assay (410 sites);

b) Additional ROs that were identified in two independent high-resolution chromatin-immunoprecipitation studies of RO licensing factors (Szilard et al., 2010; Xu et al., 2006) (52 sites);

c) Telomeric ROs that are predicted based on sequence conservation with confirmed telomeric origins (23 sites);

d) Experimentally proven false-positives were removed from the list of ROs (Müller and Nieduszynski, 2012) (4 sites).

The final list contains 482 ROs and this list considers only a single copy of rDNA (9.1 kb in size having a single RO that is duplicated ~100 times in the genome (Cherry, 2015). RO location data for four other *yeast* species were collected from published data sets for genome-wide RO positions in *Kluyveromyces lactis* (Liachko et al., 2010), *Lachancea waltii* (Rienzi et al., 2012), *Lachancea kluyveri* (Agier et al., 2013) and *Schizosaccharomyces pombe* (Hayashi et al., 2007). Even though later data sets are not of similar level of accuracy of the *S. cerevisiae* data specifically in regard to the telomeric ROs, they are strong enough to give analytical support to the RO distribution profile in *S. cerevisiae*. Genome and chromosome size information was obtained from the following sources: *K. lactis* (Dujon et al., 2004), *L. waltii* (Di Rienzi et al., 2011), *L. kluyveri* (Sherman et al., Génolevures Consortium, 2009) and *S. pombe* (Wood et al., 2002).

3.3 Results

3.3.1 Formulas for probability of DFS and TFS

In general **Model A** as being discussed in chapter 2, measures the degree of influence from RO distribution on complete genome duplication through measuring the probability of replication errors arising from DFS and TFS. We assumed no strict

temporal constraint for whole genome replication, which is biologically more plausible for cells freed from rigorous timing constraint e.g. single-celled eukaryotes or adult somatic cells. This assumption allows us to study the absolute contribution of fork stalls and its effect on the genome wide RO distribution as there is ample time for cells to use all available ROs during a cell cycle and corresponding RFs to finish their travel along the DNA strands. Under the details mentioned in chapter 2, we derived the following expressions for the probability of DFS genome wide and probability of TFS at the chromosome end as given by Eqs. ($A13$) and ($A18$).

$$P(\text{DFS in genomic bulk}) \approx \alpha \frac{N_l N_g}{{N_s}^2}(1 + R^2) \qquad (R1).$$

$$P(\text{TFS at telomeric ends}) \approx \frac{(\log 2)\, N_e}{N_s} \qquad (R2).$$

Median stalling distance N_s is the only unknown parameter while mean inter-RO distance N_l, genome length N_g, periodicity measure R that is the coefficient of variation of the inter-RO distances, sum of the distances between end-proximal RO to telomere N_e, all are directly calculated from the genome-wide RO mapping data. It is very difficult to experimentally determine the value for N_s as this is directly related to the individual RF travelling through the DNA. We obtained an experimental estimate for N_s using published DNA fibre data (Maya-Mendoza et al., 2007) which provides a rough estimate for N_s to be ~10 Mbp. Though the stall rate of RFs may vary due to impediments like repetitive regions or heterochromatin but the scale over which this variation happens has to be much smaller than the median stalling distance N_s, hence it would not significantly affect our analysis using our model formulas. Also chromosome fragile sites that are large chromosomal domains where the RFs have an increased probability of failing is thought to be regions having a paucity of effective ROs (Debatisse et al., 2012). According to Eq. ($A5$), which gives the probability of

DFS in a given inter-RO gap, this probability is directly proportional to the length of the gap in base pairs. Hence, regions like fragile site, as per our model suggests, are regions with bigger inter-RO gaps. In order to apply our formulas we first characterize the RO positions in different yeast genomes based on the RO mapping datasets (Figure 4).

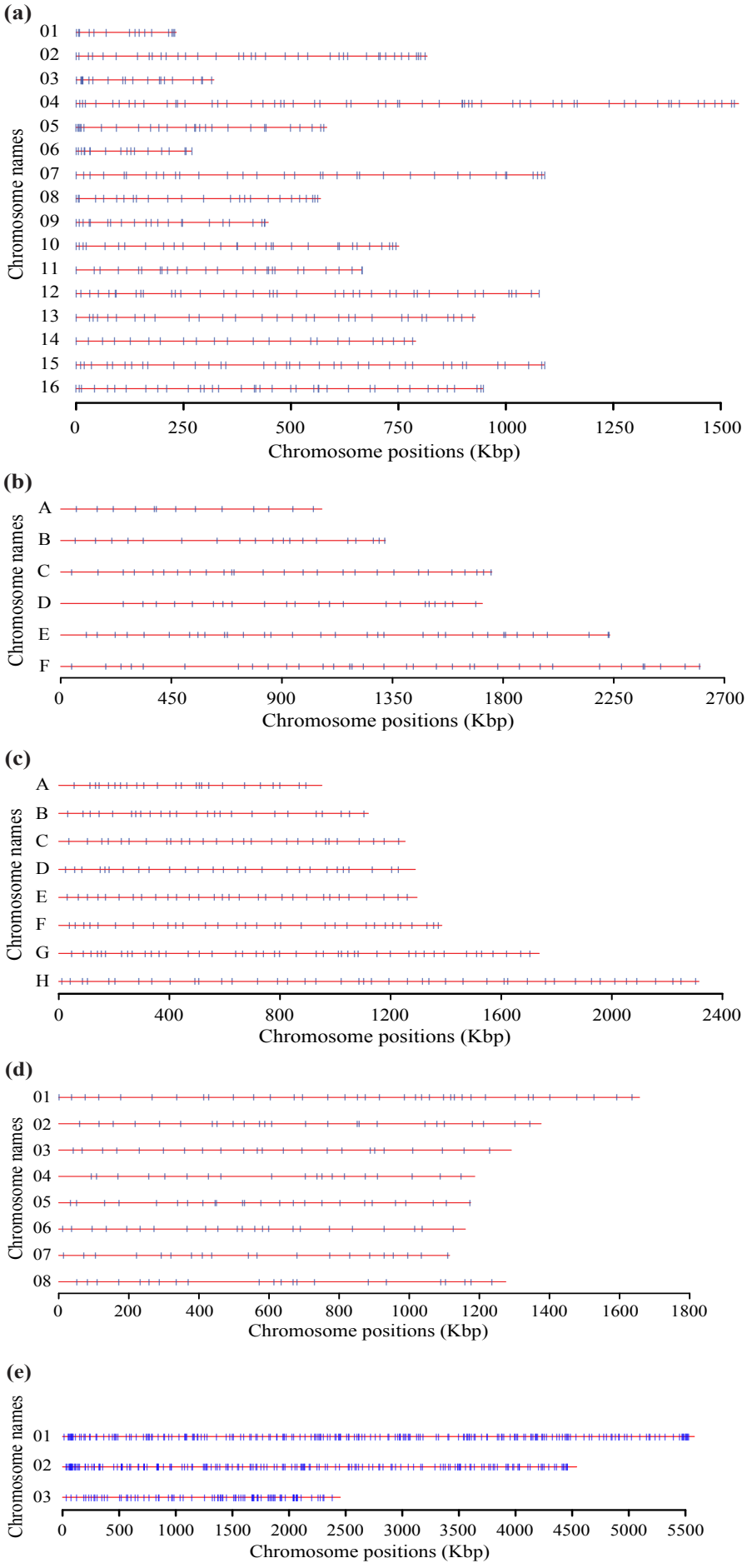

(a)
Chromosome names
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
0
250
500
750
1000
1250
1500
Chromosome positions (Kbp)
(b)
Chromosome names
A
B
C
D
E
F
0
450
900
1350
1800
2250
2700
Chromosome positions (Kbp)
(c)
Chromosome names
A
B
C
D
E
F
G
H
0
400
800
1200
1600
2000
2400
Chromosome positions (Kbp)
(d)
Chromosome names
01
02
03
04
05
06
07
08
0
200
400
600
800
1000
1200
1400
1600
1800
Chromosome positions (Kbp)
(e)
Chromosome names
01
02
03
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
Chromosome positions (Kbp)

Figure 5: RO positions are shown in a chromosome wise manner over the whole genome for (a) *Saccharomyces cerevisiae*, (b) *Kluyveromyces lactis*, (c) *Lachancea kluyveri*, (d) *Lachancea waltii*, and (e) *Schizosaccharomyces pombe*. Horizontal red lines denote individual chromosomes and short vertical blue lines denote the respective replication origin position on the chromosome.

3.3.2 Genomic distribution of ROs is non-random and biased for regularity

The genome wide inter-RO distances are supposed to be distributed in a statistical sense meaning we do not expect a completely ordered pattern. In order to check the nature of the genomic RO distribution we first calculate the inter-RO distances from the RO mapping data. We considered the mid-point of each ARS (autonomously replicating sequence) element, genomic regions that contain origins of replication i.e. ROs, as the point where the bidirectional RFs arise following the activation of respective RO. Hence the distance between mid-points of two adjacent ARS elements represents the inter-RO distance in base pairs for any given pair of adjacent ROs. Under this definition, we calculated the genome wide inter-RO distances from the RO mapping data in all five yeast species (Figure 5). We compared the genomic distribution of inter-RO distances depicted by the blue histogram, to the mean frequency of inter-RO distances in each bin in the histogram found in a simulation with randomly sampled ROs on the genome where we maintained the number of ROs and length of genome same as in the datasets. Noticeably the distinction of genomic RO spacing from random distribution is profound in *S. cerevisiae*, *K. lactis*, *L. waltii* and *L. kluyveri*. In *S. pombe*, the genomic RO distribution is closer to the random distribution in comparison to other four yeasts and this might be due to the fact that *S. pombe* does not have the sequence specific RO binding sites on the genome like the ARS elements found in *S. cerevisiae* and others.

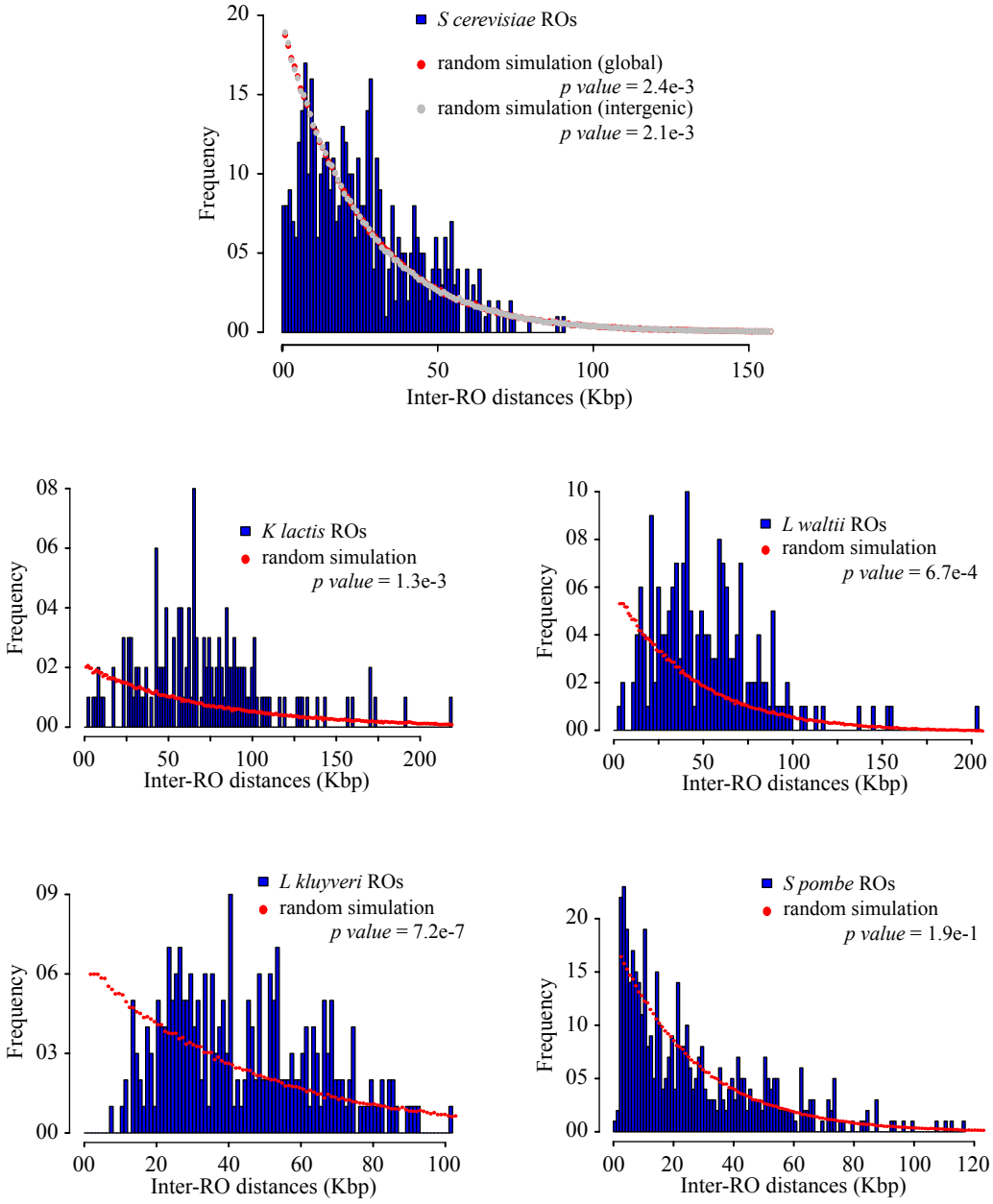


Figure 6: Histogram of genomic inter-RO distances in five different yeast species. Red-dots are representation of the bin-mean for inter-RO distances obtained from a computer simulation where equal numbers of ROs were randomly sampled over the same lengths of genome for each species respectively. For *S cerevisiae*, gray dots represent the bin-mean for inter-RO distances obtained from a random simulation with sampling of ROs being restricted only to the intergenic regions on the genome. Corresponding *p* values are calculated using a Kolmogorov-Smirnov test.

On the right hand side of the formula for the probability of DFS in chromosome bulk, we have the coefficient R, which is the coefficient of variation or the standard deviation to mean ratio for genomic inter-RO spacing. In chapter 2, we have introduced this coefficient of variation that measures the spatial dispersion of data-points in a given dataset. Standard deviation to mean ratio for complete periodic spacing of points on a line is zero meaning $R = 0$. Hence, the perfect ordering of ROs over the genome would produce $R = 0$. In our formula, the probability of DFS in the chromosome bulk is proportionally related to the value of the factor $(1 + R^2)$. Thus, setting $R = 0$ minimizes the probability of DFS and any deviation from the perfect ordering of ROs i.e. any value of R greater than zero would serve as an increase in the probability of DFS error. For this reason, if evasion of DFS error is an important issue for establishing RO positions, Eq. (1) suggests they should be more evenly distributed than expected by chance meaning the system would try to minimize the value of R as much as possible. We measured the R values genome wide as well as individual chromosomes for all five yeasts. Similar number of randomly sampled ROs on the same length of genome as in the datasets, provides a upper bound for R which is close to 1 depending on the number of ROs i.e. points sampled (Figure 3). Thus we have an scale of $R = 0$ i.e. complete order to $R \approx 1$ i.e. complete randomness to measure the spatial variation in the distribution of genomic inter-RO distances. So we compared the measured R values in chromosomes and over the whole genome to the R values obtained from simulations with randomly sampled ROs in the respective chromosomes and genomes for all yeasts (Figure 6). In all cases, R value in the data was smaller than that was obtained from random distribution. In *S cerevisiae*, R for the whole genome is 0.697 while random sampling of ROs provides a value for R as 0.999±0.046 with a very significant statistical difference given by p value 1.70×10^{-11}. This value of R does

not change much for random deletion or addition of ROs in the dataset (Figure 7c) suggesting the calculated value is robustly coping any presence of false-positives or false-negatives in the dataset. It is known that efficiency of ROs in *S cerevisiae* genome declines when transcription machinery moves through the RO (Nieduszynski et al., 2005; Snyder et al., 1988). Considering the observed low *R* value as a possible result of preferentially restricted RO placement in the intergenic regions to avoid hindrance from transcription, we performed simulation with restricting the sampling of ROs only in the intergenic regions. The intergenic simulation provided *R* value 0.997±0.046 significantly greater than the genomic *R* 0.697, thus supporting our observation that the genomic value for *R* in *S cerevisiae* genome is indeed lower than expected from a random distribution. *R* values in other four yeast species are 0.55 (*K lactis*), 0.46 (*L kluyveri*), 0.58 (*L waltii*) and 0.86 (*S pombe*), significantly lower than expected *R* in corresponding random distribution difference given by *p* values 9.99×10^{-19}, 2.51×10^{-08}, 2.48×10^{-09}, and 1.82×10^{-03} respectively (*p* values obtained from fitted normal distribution). *R* value lower than expected by chance or in other words the bias for regularity in the distribution of RO spacing signifies the pressure to avoid DFS errors is a significantly major issue in determining the distribution of RO placement in these organisms. At the same time, we observe significantly higher variation in RO spacing than expected in a periodic arrangement of ROs and this possibly reflects the biological trade off between lowering the global error rate and difficulty in placing ROs on the genome in a perfect order inside the living cell. Smaller *p* values differences to random distribution and higher *R* value in fission yeast *S pombe* could be due to fact that in this species ROs are established in a non-sequence specific manner different from other yeasts and most ROs are activated only in a small proportion of cell cycles resembling the RO features found in higher eukaryotes (Cotobal et al.,

2010; Patel et al., 2006).

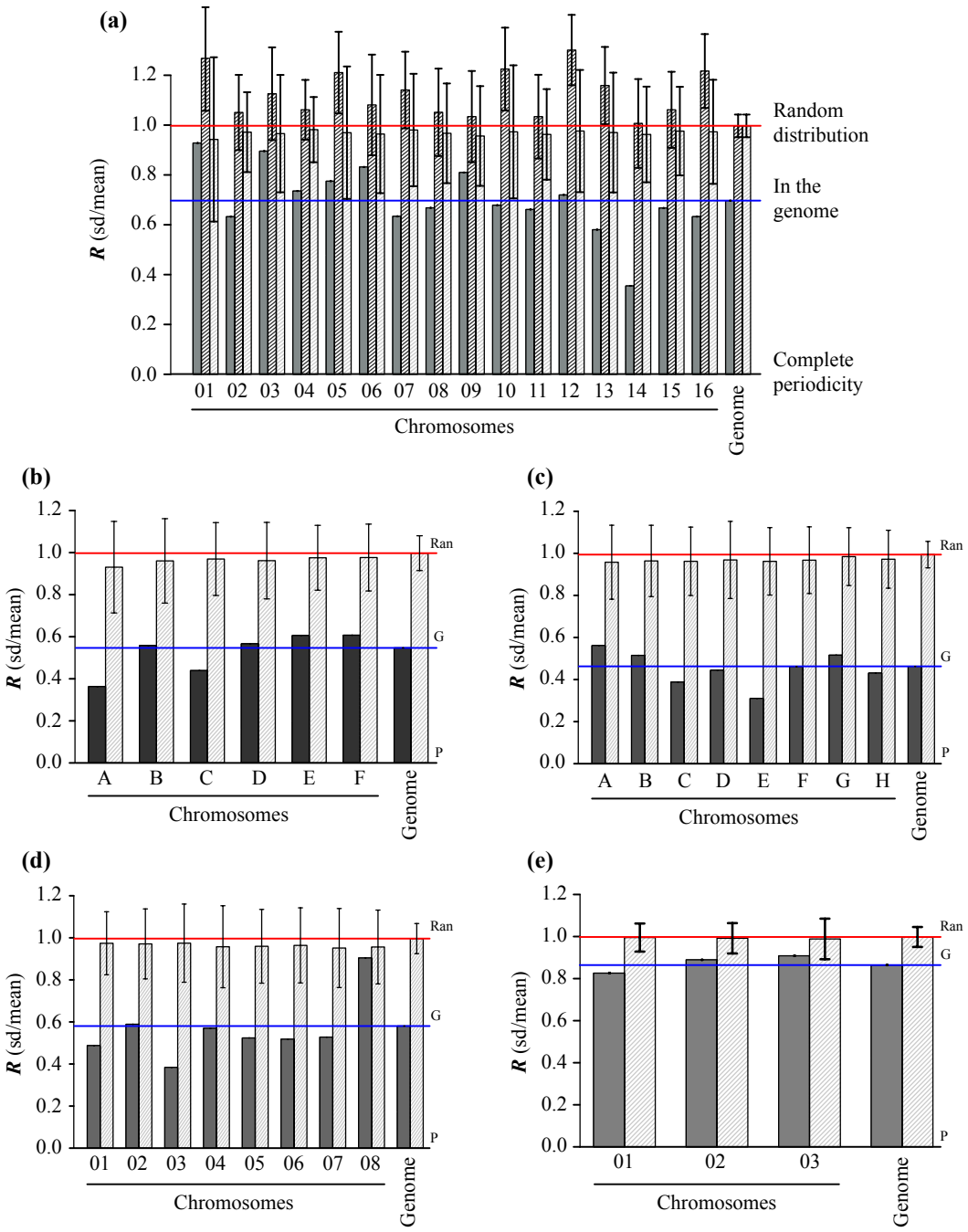


Figure 7: Coefficient of variation, *R* is shown for individual chromosomes and over the whole genome in (a) *S cerevisiae*, (b) *K lactis*, (c) *L kluyveri*, (d) *L waltii* and (e) *S pombe* alongside the *R* value obtained from simulation with randomly sampled equal number of ROs over the same length of respective chromosomes and genomes. Error bars show the standard deviation in simulation. Horizontal lines signify complete periodicity (black); genomic value (blue) and random distribution (red). In *S cerevisiae*, bars in the middle show the intergenic simulation results.

3.3.3 Largest inter-RO distance is smaller than expected by chance

Probability of a DFS inside a given pair of adjacent ROs is proportional to the square of the distance between the two ROs as given by the Eq. (*A*5), which suggests increased probability of DFS as the distance between the respective RO pair increases. Hence, the largest inter-RO gap will be the most error prone site in the genome as the probability of DFS is maximum in this gap and for this reason the largest inter-RO gap requires special attention. The number of licensed ROs in a genome is large and we already saw in the previous section that the *R* value in the genome is much less than one meaning the RO distribution is not aperiodic enough to be random but at the same time the distribution is far away from proper periodicity. Thus the degree of aperiodicity available in the genomic RO distribution with the number of ROs licensed carries the rare possibility for large inter-RO gaps that may lie far away from the standard deviation of the distribution. We checked the largest inter-RO gaps in all the datasets and compared it to the largest separation between adjacent RO pairs found in simulation where same number of ROs was randomly sampled on respective genomes (Figure 7b). In *S cerevisiae*, the maximum inter-RO gap is 90.1 Kbp while the average maximum gap from the simulation was 169±31 Kbp and the genomic value had a very significant difference with the random simulation measured by the *p* value 4.05×10^{-07} (obtained from a fitted Gumbel extreme value distribution). Similarly, when we restricted the random sampling of ROs only in the intergenic regions the simulation average for maximum gap was 182±34 Kbp and the genomic value had a difference with the simulation measured by *p* value 6.13×10^{-09}. In other four yeasts the genomic maximum inter-RO gaps and respective simulation average were 219 Kbp and 395±78 Kbp (*K lactis*); 102 Kbp and 267±55 Kbp (*L kluyveri*); 203 Kbp and 299±60 Kbp (*L waltii*); 116 Kbp and 182±34 Kbp (*S pombe*). The significance of the difference

between genomic value and simulation results are given by p values 2.37×10^{-04} (*K lactis*); 8.78×10^{-16} (*L kluyveri*); 1.08×10^{-02} (*L waltii*); 9.52×10^{-04} (*S pombe*).

The fact that the largest inter-RO gaps in all yeasts are much smaller than expected by chance again emphasizes our observation from R value analysis that the pressure for avoiding DFS errors is a major player in determining the placement of ROs in these genomes. Due to the importance of the largest inter-RO separation, we wrote down the formula to calculate probability of DFS in this gap denoted by $N_{\max}$ in **Model A** given by Eq. (A14) as the following:

$$P(\text{DFS in } N_{\max}) = \alpha\left(\frac{N_{\max}}{N_s}\right)^2 \qquad (R3).$$

Combining Eq. (R1) and (R3), we checked the relative contribution of the biggest inter-RO gap to the over all genome wide probability of DFSs using the following expression:

$$\frac{P(\text{DFS in } N_{\max})}{P(\text{DFS in genomic bulk})} \approx \frac{(N_{\max})^2}{N_l N_g(1+R^2)} \qquad (R4).$$

The largest inter-RO separation in S cerevisiae is 90.1 Kbp which roughly 0.7% of the genome and this gap represents 1.8% of the genome wide probability for DFSs while the mean obtained from the randomly sampled ROs on this genome 169 Kbp represents 4.6% of the over all DFSs probability. The similar is true for all other yeast species and hence the maximum gaps in the genome being limited compared to the random simulation is farther enhancing to the motive to arrange the RO distribution in a manner to minimize the chances of DFSs. There have been experimental efforts in *S cerevisiae* to study artificially induced large RO less regions where deletion of five ROs between ARS304 and ARS313 in chromosome 3 created 160 Kbp RO less region

(Theis et al., 2010; construct 5ORIΔ). The observed per cell cycle chromosome loss rate of construct 5ORIΔ was ~ 9×10^{-05} while the comparable test chromosome 0ORIΔ-ΔR with no deletion of ROs showed chromosome loss rate of ~ 3×10^{-05} implying an increase of ~6×10^{-05} in the rate of chromosome loss due to the artificial RO less region. Our Eq. (*R*3) with inter-RO distance set to the value of 160 Kbp and considering N_s as 10 Mbp, provides a value for the probability of DFS in this gap as 6.1×10^{-05} that is remarkably similar to the observed chromosome loss rate in the above mentioned experiment, which significantly strengthens our theory and also serves as a support for the estimation of N_s as 10 Mbp.

3.3.4 Telomeric ends are much smaller than inter-RO distances in the genome

Telomeric ends in the linear chromosome constitutes an important issue for complete genome replication as the remaining DNA after the end-proximal RO needs to be replicated by a single RF either originated at the end-proximal RO or from else where that has passively replicated the last RO. Hence, there are no more ROs to compensate if this lone RF stalls (TFS, Figure 2) before reaching to the very end of the chromosome. Now, we would expect this end-proximal RO to telomeric end distance to be significantly smaller in comparison to the inter-RO distances in the bulk of the chromosomes if the pressure to avoid replication failures arising from RF-stalling is indeed playing a role in determining the placement of ROs on the genome. To verify this, we checked the position of end-proximal ROs in *S cerevisiae*. The average of the distance between end-proximal RO to the telomeric end from all 16 chromosomes is 404±273 base pairs, which is two orders of magnitude smaller than the value 26±18 Kbp for inter-RO distances present in the chromosomal bulk (Figure 7d). Among all the 32 chromosome ends the maximum end-proximal RO to telomeric end distance is 730 base pairs only. In effect, this is another strong evidence to support the argument

the RO placement in yeast genome is under strong influence from the need to avoid replication errors from RF-stalling events.

3.3.5 RO distribution in yeasts maintain a low replication failure rate

We can directly calculate the global probability of DFS error over the whole genome using Eq. (R1) after we have already calculated the values of R in the genomes. In Eq. (R1), probability of DFS error in genomic bulk is proportional to the factor $(1 + R^2)$ = 1.49 (with R = 0.697 in *S. cerevisiae* genome) which is exactly in the middle between complete periodicity (R = 0, hence $1 + R^2 = 1$) and complete randomness (R = 1, hence $1 + R^2 = 2$). In *S. cerevisiae*, the value for N_g is 12.1 Mbp and from the dataset we have N_l as 25.9 Kbp. Even though, direct experimental estimate for median stalling distance N_s is hard to obtain but alongside the suggestion extracted from DNA fibre experiment in human cell lines for N_s as ~10 Mbp (Maya-Mendoza et al., 2007), we already presented support for this value of N_s from our model during chromosome loss rate analysis due to large RO less region in *S cerevisiae*. Using these values of N_g, N_l, N_s and R as 0.697 in Eq. (R1) provides the probability of DFS in *S cerevisiae* genome to be 0.11%. In *S cerevisiae*, individual chromosomes missegregate in ~2×10^{-05} of all cell divisions (Spencer et al., 1990; Strome et al., 2008) and thus any of the 16 chromosome would have a missegregation rate of $16\times2\times10^{-05}$ or 0.032% which is within a factor of three to our calculated probability for DFSs. Due to the fact that both chromosome missegregation and replication errors serve for chromosomal instability, this magnitude equivalence is not unexpected. In other four yeasts, with N_s as 10 Mbp we calculated the genome wide probability of DFSs: 0.24% (*K lactis*); 0.14% (*L kluyveri*); 0.17% (*L waltii*); 0.14% (*S pombe*) which are all very close. We have compared these probabilities to the simulation with randomly sampled ROs (Figure 7a) and in all organisms the probability of DFS is much smaller than expected by chance. This

implies the similar genome sizes in these organisms are keeping a similar tolerable range of probability of DFSs quite small by the virtue of pressure to avoid DFSs influencing RO distribution. This small probability being of similar magnitude to chromosome missegregation rate suggests indeed the pressure to avoid DFSs is a well-reflected issue in maintaining global replication fidelity in these organisms.

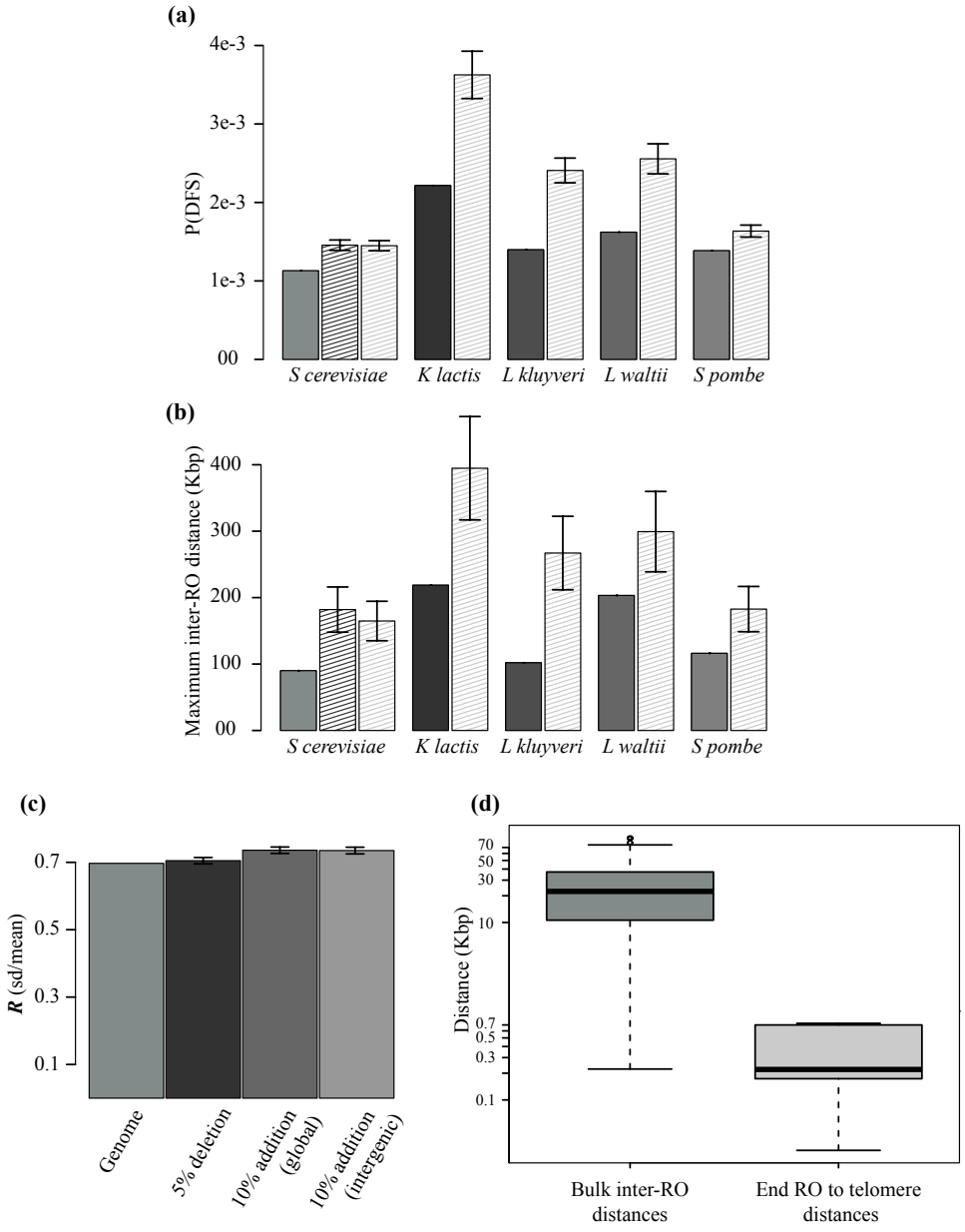


Figure 8: a) Genome wide probability of DFSs, calculated using our model formula, is shown beside the random simulation results with sampling same number of ROs on the genome. In *S cerevisiae*, the middle bar shows the probability obtained from

simulation with restricting the sampling only in intergenic regions. b) The measured maximum inter-RO distances in five yeasts are shown beside the simulation results. c) Variation of R value in S cerevisiae genome is shown due to random addition and deletion of ROs. d) The distances between end-proximal ROs to the telomeric ends in 32 chromosomal ends of S cerevisiae is compared with inter-RO distances in the genomic bulk.

3.3.5 Spontaneous stalling distance is ~10 Mbp

The length of DNA that a single RF can replicate in absence of any checkpoint regulation and replicative stress before it stalls is the spontaneous stalling distance for RFs. We defined the average distance a RF will travel before stalling by N_s - the median of the spontaneous stalling distance. Alternatively, this can also be characterized by the per nucleotide stall rate. For replication faithfulness, by necessity this spontaneous stall distance has to be much bigger than individual replicons and thus it becomes difficult to determine the value of this experimentally and the indirect estimates available so far are not well characterized. Due to the paramount importance of RF stalling distance in the dynamics of replication machinery and their efficiency and therefore to the overall replication fidelity, it is very much urgent that we study and shed light in this regard. In our model, we have only one unknown parameter, which is N_s, thus up to an extent validation of our model depends on the verification for the value of N_s. We derived an indirect experimental estimate for N_s from DNA fibre studies as ~10 Mbp (stall rate ~6×10^{-08} /base pairs) and showed a support for this in the analysis of chromosome loss due to large inter-RO gap. Nevertheless in order to verify this estimate, we sought to use our formulas for probability of DFSs in the genome and probability of TFSs in the telomeric end. To keep the integrity of replication fidelity, overall probability of RF-stalling events should be balanced through out the genome. Hence intuitively the TFS probability at the end should be

similar to the DFS probability in the chromosomal bulk by making this very much biologically plausible assumption we derived a formula for N_s in **Model A (**Eq. (*A*19)):

$$N_s \approx \left(\frac{\log 2}{2}\right)\frac{N_l N_g}{N_e}(1+R^2) \qquad (R5).$$

Using the *S cerevisiae* values for N_l, N_g, R and N_e, which is the sum of all 32 telomeric ends (12696 base pairs), we get $N_s \approx 12.7$ Mbp (stall rate 5.4×10^{-08} /base pairs) which is in very good agreement with the estimation from the DNA fibre study (Maya-Mendoza et al., 2007). This remarkable convergence in the numbers provides strong confidence to our theory as well as reinforces the estimated N_s value to be factual and thus it has great influence in the determination of RO distribution on the genome too. In the study of artificially induced large RO less region in *S cerevisiae* chromosome, beside creating RO less region of 160 Kbp inside the chromosome body (construct 5ORIΔ) there was also the similar RO less region created at the end of the chromosome (construct 0ORIΔ-ΔR) (Theis et al., 2010). Due to the RO less region being at the end, the chromosome loss rate increased from $\sim9\times10^{-05}$ to $\sim210\times10^{-05}$ in agreement with our suggestion that unidirectional replication of telomeric end increases the replication failure rate. Applying Eq. (*R*2) with end distance as 160 Kbp provides a chromosome loss rate of $\sim1000\times10^{-05}$ in every cell division almost 5 times larger than observed in the experimental study. This could possibly imply the presence of additional mechanisms at the telomeric ends to ensure complete replication of the whole genome.

3.4 Discussion

If an RF stalls irreversibly on the genome, replication can still be completed by an RF travelling in the opposite direction or by means of activating otherwise dormant ROs. However if both RFs travelling in opposite direction irreversibly stalls within a region bounded by a pair of adjacent ROs where there is no more ROs to do the recovery, the remaining unreplicated DNA inside the two stalled RFs renders severe consequences for the cell. The same is true with the stalling of the lone RF at the very end of the linear eukaryotic chromosomes which is replicated in unidirectional manner by a single RF, the DNA would remain unreplicated if the last RF stalls before reaching to the chromosomal end. Eq. ($R1$) directly relates the genome length N_g, genomic mean of inter-RO distances N_l, median stalling distance N_s with the coefficient of variation R of the RO distribution, which is a spontaneous measure for the degree of regularity in genomic RO spacing, in order to calculate the genome wide probability of DFSs. The equation suggests if the ROs are periodically placed, the probability is minimum and we have shown that the genomic RO distribution in all five yeasts is much more regular than expected if ROs were randomly sampled on the genome. Eq. ($R3$) shows the probability of DFS within an inter-RO gap is proportional to the square of the gap size and therefore the maximum gap in the genome should be limited in order to avoid DFS in that region. We have shown from the data that in yeasts this maximum inter-RO gap is significantly smaller than expected by chance. Moreover, due to the special unidirectional replication at the chromosome end, to avoid replication failure arising from stalling of the lone RF at the end, there are ROs placed very close to the telomeric end and we have shown this end-proximal RO to telomeric distance is much smaller than the inter-RO distances at the chromosome body.

All together 1) bias for regularity in RO placement, 2) limited maximum inter-RO separation and 3) very small distance at the end to be replicated by the lone RF, very significantly supports the idea that RO placement in yeasts is under strong influence from the pressure to avoid replication impairment from irreversible RF stalling events. Previous studies have shown that the avoidance of DFSs depends on the actual number of ROs licensed rather than the issue of whether they have used efficiently or remained dormant (Blow and Ge, 2009) and also if the RO efficiency falls low, ROs are clustered together to decrease the effect on replication time rather than by increasing the periodicity in the RO distribution (Karschau et al., 2012). Therefore, the degree of periodicity we observed in the genomic RO distributions is most conceivably due to the pressure to avoid stalling of RFs.

By equating the expressions for probability of DFS and TFS, in our model we derived the formula for spontaneous RF stalling distance N_s, which we obtained as 12.7 Mbp using *S cerevisiae* values for necessary variables in the formula i.e. N_g, N_l, N_e and R. This value for N_s is in very good agreement to the previous experimental estimate of ~10 Mbp from DNA fibre studies. Using N_s as 10 Mbp in Eq. ($R1$) and ($R2$), we get the probabilities for DFS and TFS in S cerevisiae as 0.11% and 0.09% respectively. The similitude in these two numbers as well as their closeness to the spontaneous individual chromosome missegregation rate of 0.032% (which also contributes to genome instability alongside RF stalls) greatly signifies the biological balance in genome wide protection against the consequences of irreversible RF stalling and moreover suggests that our model is capturing well the systemic features of genome replication in the organisms under consideration. Despite this there are possible competing issues to influence the RO distribution such as 1) transcription through the ROs in *S cerevisiae* decreases the efficiency of ROs (Nieduszynski et al., 2005), for

which we compared the data with randomly sampled ROs only in the intergenic regions and we saw overall data features were similarly different from the results obtained from random simulations where we sampled ROs over the whole genome without the intergenic restriction; 2) the variation in spontaneous stalling distance N_s genome wide due to the structural complexity and impediments on the DNA strands might influence the RO distribution, but we showed the value of N_s is approximately same as the genome size and is much larger than any inter-RO distance in these organisms and hence the conceivable variation in N_s would have negligible impact on the RO distribution in these range of genome length. Therefore, we strongly suggest that in the organisms under consideration i.e. yeasts, RO distribution is under very significant influence from the need of avoiding RF stalling in order to confirm complete replication of the genome.

Eq. ($R1$) provides a proportional relationship between the probability of DFS on the genome and the length of genome N_g. With N_g as around 10 Mbp (in yeasts) application of the equation provides a very small probability of DFS errors, which is actively kept under tolerable range (~1 in a thousand cell division) by having a degree of regularity in RO distribution ($R < 1$). However, organisms with much bigger genomes such as human i.e. diploid genome length are ~6000 Mbp, shows a very different scenario when we applied our equation to this genome. Due to the proportionality of N_g to the probability of DFSs, the probability in this large genome is very high and this high probability can be reduced by decreasing the inter-RO distance N_l but to achieve the level observed in yeasts, N_l needs to be almost ~50 base pairs which is biologically implausible due to the fact of MCM2-7 foot-print being ~70 base pairs long (Evrin et al., 2009; Remus et al., 2009). Moreover, the mean for genome wide inter-RO distances N_l in human cell lines is around 10-30 Kbp, and the value for

R is around 1.5 to 3 (Besnard et al., 2012; Picard et al., 2014). Hence, the probability of DFSs genome wide in longer genomes would be very high according to our model and this high probability for DFSs renders for a need to have additional mechanisms to manage the consequences of DFS errors in these organisms. Therefore, we finish this chapter by making the suggestion that the higher eukaryotes like human has extra safeguards to deal with the DFS error in a efficient way in order to maintain the replication fidelity.

Chapter 4

Inevitable errors require containment during replication in higher eukaryotes

4.1 Brief introduction

In previous chapters we have discussed in detail how double fork stalls (DFSs) are a key challenge for complete genome duplication in eukaryotic cells. We have shown in chapter 3 that in yeasts, the rate of DFSs is maintained below a tolerable limit by means of evenly distributed RO positions over the ~10 Mbp genome. However, due to the much larger genome lengths along in higher eukaryotes, with the presence of more complex checkpoint mechanisms during replication, the process of maintaining replication fidelity is more complex. The implication of DFSs in this context is much greater as larger genome lengths increase the error probability too. Moreover, due to various studies showing the association of cancer and other diseases with insults to the biological processes involved in DFSs, the study of the challenge from DFSs to complete replication and its resolution in higher eukaryotes is of high biomedical importance (Ghosal and Chen, 2013; Macheret and Halazonetis, 2015; Mazouzi et al., 2014). For this purpose, in this chapter we are going to apply our model to different eukaryotes with genome lengths spanning Megabases to Gigabases i.e. yeasts (diploid genome ~20 Mbp), *Arabidopsis* and *Drosophila* (diploid genome ~250 Mbp) and human cell lines (diploid genome ~6000 Mbp). In chapter 3, for the restricted application of the model to yeasts, the assumption that all the genomic inter-RO distances are much smaller than RF-stalling distance N_s was valid because of the smaller genome length. Nonetheless, as the genome length is much larger in higher eukaryotes the possibility of variation in inter-RO distances is high and thus in order to

make the model more suitable for higher eukaryotes, we removed this assumption in the extended model which has been presented as '**Model B**' in chapter 2. Application of this extended model on the RO mapping data from different eukaryotes with different genome lengths yields some clear predictions, which are very much in agreement with the data. As the genome length changes from Megabases to Gigabases, the bias for regularity in RO distribution is lost; in bigger genomes, larger inter-RO distances contribute most towards the error rate, but the largest inter-RO gap is constrained by N_s, and in larger genomes the error becomes increasingly inevitable but the occurrences of errors are low in number and are distributed as Poisson. We are going to discuss these in depth, in the results section below.

4.2 RO distribution data in different eukaryotes

Most of the data available of ROs in plants and higher eukaryotes are more focused on genomic density of ROs rather than the locations (Mahbubani et al., 1997; Wong et al., 2011). However, to calculate the probability of error in a given genome our model requires the genome wide RO positions as input. Hence, we used only the datasets describing genome wide RO positions for different eukaryotes in consideration.

- *Saccharomyces cerevisiae* ROs were obtained from the highly curated OriDB (Siow et al., 2012) with selection criteria discussed in chapter 3. For additional validation, we included another yeast species in this chapter: *Schizosaccharomyces pombe* (Hayashi et al., 2007).
- RO distribution data were also obtained for the following multicellular organisms: *Arabidopsis thaliana* (Costas et al., 2011), *Drosophila melanogaster* (Cayrou et al., 2011) and human.

- Human data for the four cell lines IMR90, HeLa, hESC and iPSC were derived from (Besnard et al., 2012) and different datasets for IMR90, HeLa and K562 cell lines were obtained from (Picard et al., 2014). Since, Picard et al. used more modern techniques, in particular for peak detection, it might be considered as a more reliable dataset while the comparison with the Besnard et al. dataset is an useful assessment to the experimental uncertainties and variability in the data.
- Due to the limitation in sequencing the centromeric region, we excluded the inter-RO distance corresponding to the centromeric region in each chromosome from the analysis in all the organisms considered.

4.3 Results

4.3.1 The '**central equation**' for determining replication errors

In '**Model B**' we have derived a more general formula for the probability of DFSs in a given genome with the consideration that due to the very large genome length in higher eukaryotes there is an occasional possibility of finding arbitrarily large inter-RO distances, hence the variation in inter-RO distance should not considered restricted in larger genomes as we assumed for yeasts, rather the new formula is well suited for broad variation in inter-RO distances. Eq. ($B10$) in chapter 2, gives the formula for genome wide probability of DFSs in '**Model B**' as the following:

$$\text{Prob(zero DFS)} = \exp\left(-\frac{\log(2)\,N_g}{N_s} + \sum_k \log\left(1 + \frac{\log(2)\,N_k}{N_s}\right)\right) \qquad (R6).$$

where N_g is the genome size, N_s is the median fork stalling distance and N_k is the distance between any adjacent pair of ROs, k and $k+1$.

As the probability of a DFS in any given inter-RO distance is small, we have shown in chapter 2 under '**Model B**' that the statistics of DFSs are Poisson to a very high level

of accuracy. We, here remind the reader from chapter 2 that the probability of zero events in a Poisson distribution is given by the following:

$$\text{Prob}(\text{zero events}) = \exp(-\lambda), \qquad (R7).$$

where λ is the single Poisson parameter that describes the mean as well as the variance of the Poisson distribution. From Eq. (*R*6) and (*R*7), it is clear that the probability of no DFSs genome wide has the form $\exp(-\lambda)$. Thus, a great deal of information concerning the probabilities of DFSs for given genome can be obtained from the single parameter λ. Combining Eq. (*R*6) and (*R*7), we can obtain the direct formula to calculate λ in the genomic distribution of inter-RO distances as presented below:

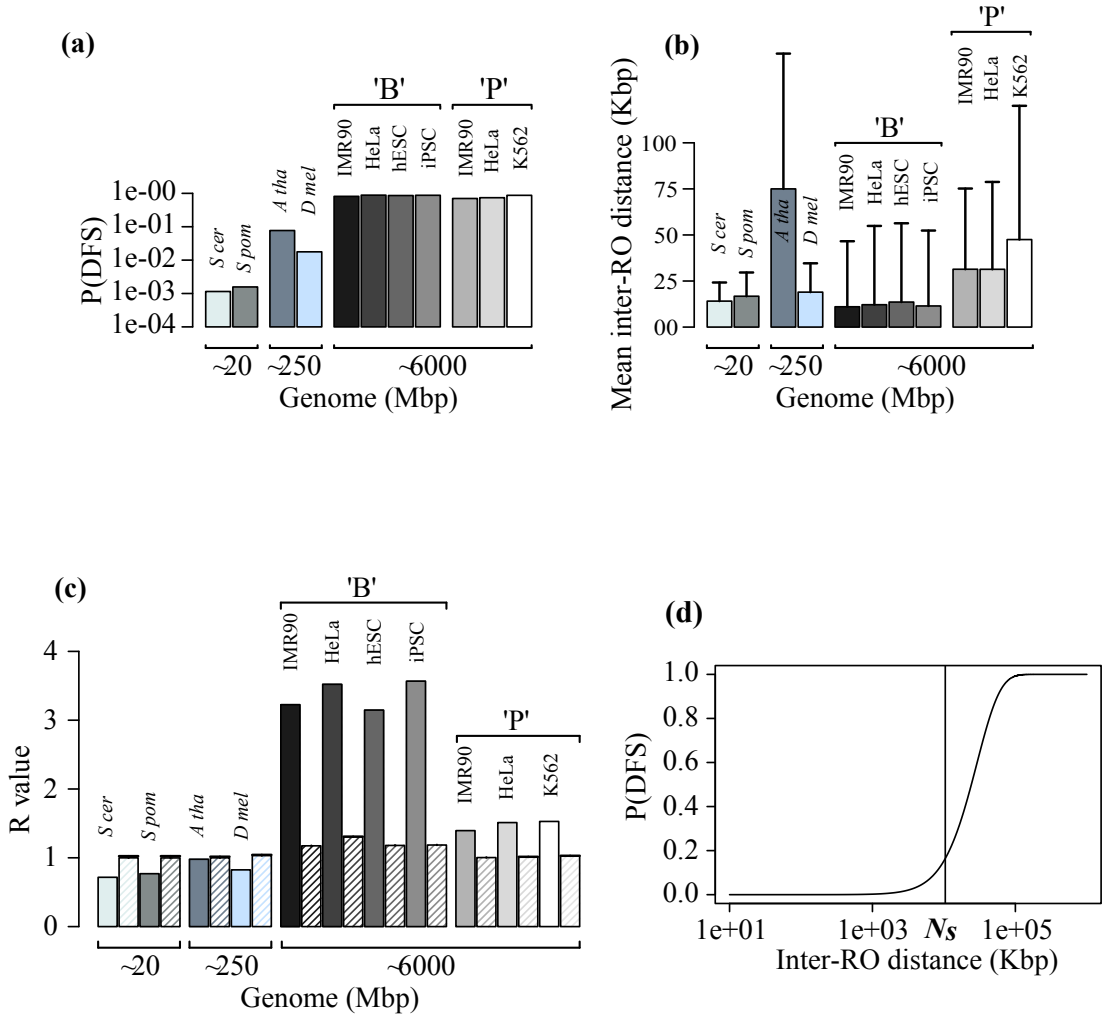
$$\lambda = \log(2)\frac{N_g}{N_s} - \sum_{i=1}^{K} \log\left(1 + \log(2)\frac{N_i}{N_s}\right) \qquad (R8)$$

This expression for λ contains a single unknown parameter N_s – i.e. the median stalling distance for RFs, or in other words the number of replicated bases along the DNA beyond which 50% of the RFs irreversibly stall. This is inversely proportional to the very small probability of stalling per nucleotide and we have derived the value for N_s in chapter 3 as 12.7 Mbp which is very much in agreement with the experimental estimate available for this value as ~10 Mbp (Maya-Mendoza et al., 2007). On the right-hand-side of Eq. (*R*8), we can identify the two distinct contributions of the genome length (first term) and of the RO distribution (second term). Genome length determines a baseline probability of DFSs that can be lowered by increasing the number of ROs and/or changing their distribution along the genome: indeed, as we have shown in previous chapter, for a given number of ROs, equally distributing them across the genome is the optimal arrangement to minimize the probability of DFSs. This establishes a hierarchy of contributions to the probability of DFSs, with genome length being the most important factor, followed by RO number and then RO

distribution. In organisms with relatively small genomes, such as yeasts (~10 Mbp), an average density of 1 RO per ~20 Kbp allows the maintenance of very small probabilities of genome-wide DFSs. Application of Eq. (*R*8) to the yeast datasets gives values around 10^{-3} for the probability of one or more DFSs, consistent with our previous analysis. With the increase in genome size from around 10 Mbp (in yeasts) to around 10 Gbp (in human), Eq. (*R*8) shows that the probability of DFSs increases by approximately two orders of magnitude, to more than 0.5 for human genomes (Figure 9a). This huge increase in error rate occurs despite essentially no shift in the mean inter-RO distance (Figure 9b). For this reason it is absolutely necessary for these organisms to have molecular mechanisms able to repair DFSs.

4.3.2 Bias for evenly spaced ROs is progressively lost in larger genomes

The regularity of the RO distribution can be assessed by computing the coefficient of variation of the inter-RO distances, denoted by *R*, which is the ratio of their standard deviation to their mean. In chapter 2, we have shown, for a perfectly uniform distribution of equally spaced ROs, *R* is equal to 0 while numerical analysis showed that when ROs are randomly distributed on the genome, the value of *R* is very close to 1 in a way that depends on the number of ROs considered (Figure 3). In the yeast genomes (diploid genome sizes ~20 Mbp), we showed in chapter 3 that the RO distributions are strongly biased towards uniform spacing with values of *R* ranging from 0.46 to 0.86 and specifically for the two yeasts we considered in this chapter the values of R are 0.69 (*Saccharomyces cerevisiae*) and 0.86 (*Schizosaccharomyces pombe*) (Figure 9c). The probability of DFSs is very small in yeasts due to the small genome size, and optimization of the RO positions by lowering *R* reduces this even further.


(a)
P(DFS)
1e-00
1e-01
1e-02
1e-03
1e-04
S cer
S pom
A tha
D mel
IMR90
HeLa
hESC
iPSC
IMR90
HeLa
K562
'B'
'P'
~20
~250
~6000
Genome (Mbp)
(b)
Mean inter-RO distance (Kbp)
100
75
50
25
00
'B'
'P'
Genome (Mbp)
(c)
R value
4
3
2
1
0
Genome (Mbp)
(d)
P(DFS)
1.0
0.8
0.6
0.4
0.2
0.0
1e+01
1e+03
Ns
1e+05
Inter-RO distance (Kbp)

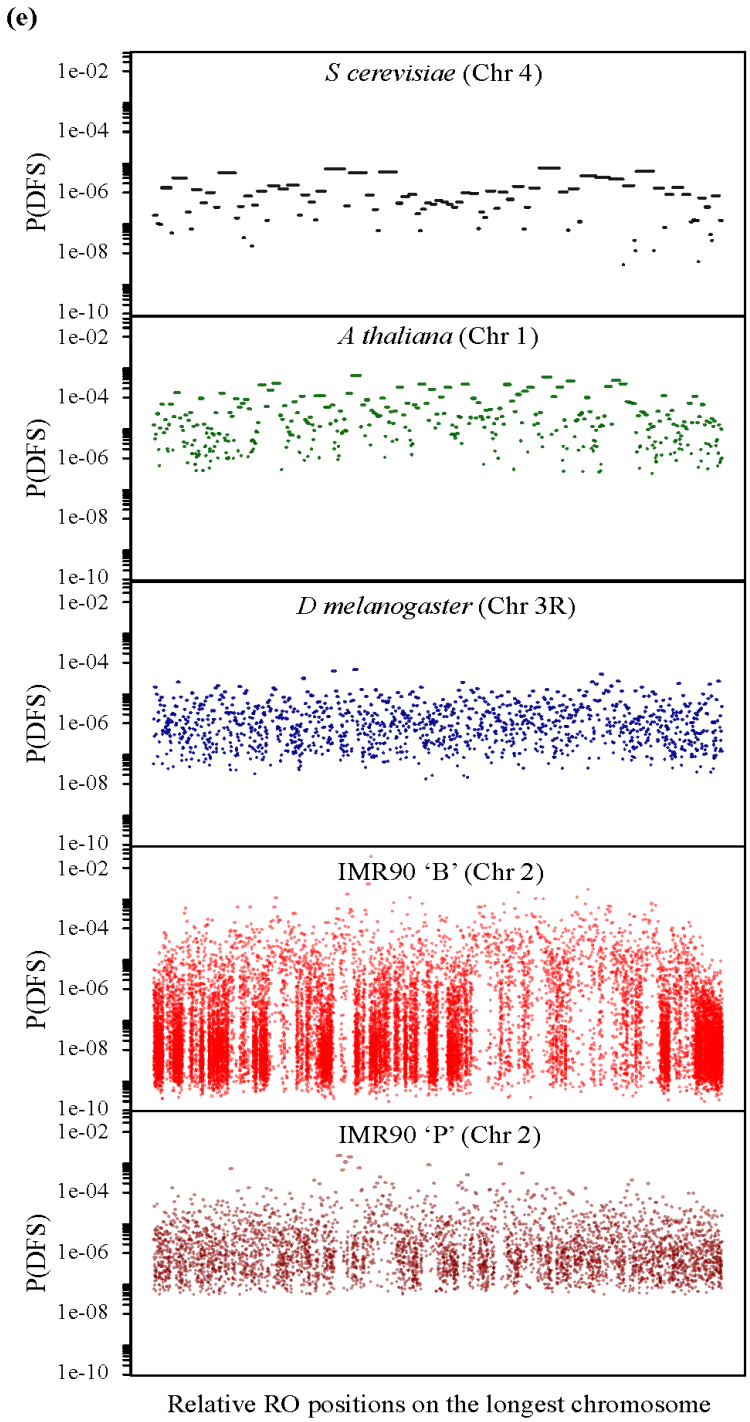


Figure 9: a) Probability of DFSs for various eukaryotic genomes from the RO mapping datasets; b) Measured mean inter-RO distances in the genomes and bar represents the positive standard deviation in the inter-RO distances data; c) Computed *R*-values in the genomes; note, the dashed bars represent simulated *R*-values for virtual genomes of the same length and RO density, but assuming ROs to be randomly distributed. d) The probability of a DFS, denoted P(DFS), is plotted as a function of increasing inter-RO distances. The estimated median fork-stalling distance, N_s (10 Mbp), is highlighted on the *x*-axis; e) The calculated probability of a DFS inside each inter-RO distance plotted against normalized chromosomal lengths for the largest chromosomes in budding yeast, *Drosophila*, *Arabidopsis* and the IMR90 cell-line from two human datasets ('B' for Besnard et al., and 'P' for Picard et al., datasets).

However, as discussed above, organisms with larger genomes have a significantly higher probability of DFS events, which results in the need for additional molecular mechanisms to cope with the consequences (Moreno et al., 2016) and the presence of such mechanisms means there is little to be gained in uniformly ordering ROs on the genomes. Thus, our expectation is that R should be significantly larger in organisms with larger genomes as the pressure to maintain regularity in RO spacing is no more present in these genomes as well as the large genome with large number of ROs provides the scope for much broader variation in inter-RO distances providing the possibility for wider standard deviation and hence the ratio of standard deviation to mean would be higher too. Statistical analysis of the available data confirms this expectation (Figure 9c). *Arabidopsis* and *Drosophila* (diploid genome sizes ~250 Mbp) have values of R around unity (i.e. approximating a random distribution). Particularly striking is the fact that in human genomes (~6000 Mbp), the values of R are *significantly larger* than unity, indicating that ROs are not spaced purely randomly and that both the number and size of large inter-RO distances is significantly greater than expected by chance. These large inter-RO distances has important consequences due to the relationship between length of the inter-RO gap to the probability of a DFS occurring in that particular gap. The probability of a DFS in a given inter-RO distance increases with the size of the inter-RO distance according to Eq. (6) as presented in the general model in chapter 2. This relationship between length of inter-RO distance and the probability of DFS in the gap is plotted in Figure 9d. The probability has a strongly non-linear form i.e. increases as the square of the lengths of the inter-RO distances which are much less than the median stalling distance N_s, and saturates at unity for lengths significantly greater than N_s. Figure 9e provides a graphical representation that highlights the shift in variation of inter-RO distances, or equivalently the probability of

DFS per inter-RO distance, by plotting the predicted probability of DFSs across the largest chromosome of different organisms that are considered for our analysis in this chapter. It is apparent that the variation in probability of error increases by approximately one order of magnitude from yeast to *Drosophila*, and then again by approximately one order of magnitude from *Drosophila* to human.

4.3.3 Large inter-RO distances contribute most to error but are bounded by the fork stalling distance in human genome

In order for the consistency to our analysis of the values of R, we expect the largest inter-RO distance in the genome to be significantly distinct in size and also in their comparison to the random simulation results in different diploid genomes: ~20 Mbp, ~250 Mbp and ~6 Gbp (represented by yeasts, *Drosophila*/*Arabidopsis* and human respectively), with significantly larger inter-RO distances appearing in those genomes with R larger than unity. As seen in Figure 10a, this is exactly what is observed, with the largest inter-RO distances being ~60 Kbp in yeasts (~120 Kbp expected for a random distribution), 151 Kbp in *Drosophila* (207 Kbp expected if random), 773 Kbp in *Arabidopsis* (663 Kbp expected if random) and ~5 Mbp in human (~300 Kbp expected if random). This can also be observed by the significant increase in outliers in the box plots of inter-RO distances for the different organisms considered (Figure 10b). As is clear from Figure 9d, the probability of a DFS in a given inter-RO distance increases sharply as the length of the inter-RO distance approaches the median stalling distance N_s. In order to avoid almost inevitable errors arising from single inter-RO gaps, we would expect the largest inter-RO distance in the genome to be bounded by N_s, and this is indeed what is observed in the data.

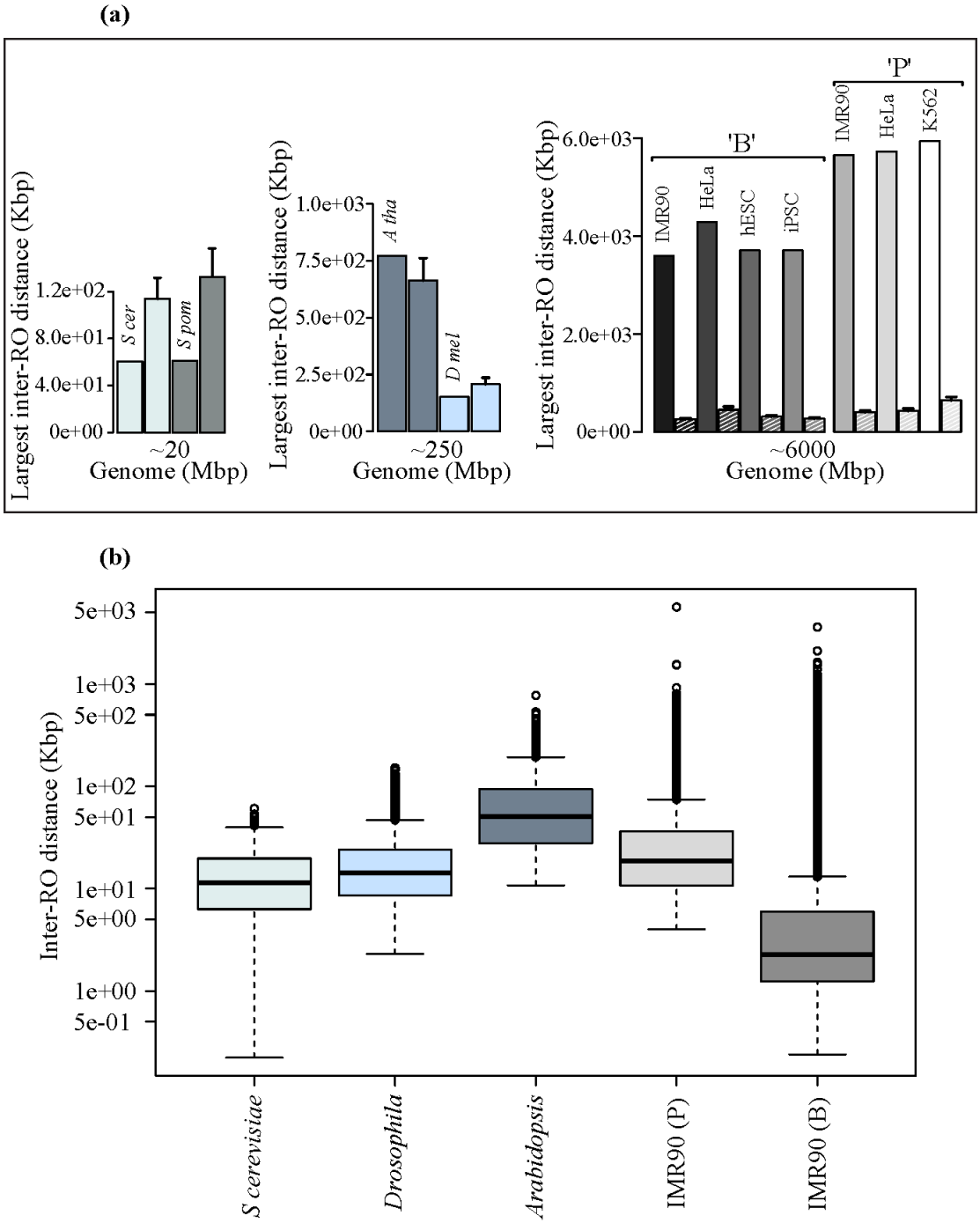


Figure 10: a) Measured lengths of the largest inter-RO distances are shown in each dataset alongside the dashed bars showing the value obtained for virtual genomes of the same length and RO density, but assuming ROs to be randomly distributed; b) The distribution of genome-wide inter-RO distances plotted in boxplot format for budding yeast, *Drosophila*, *Arabidopsis* and the IMR90 cell-line from two human datasets ('B' for Besnard et al., and 'P' for Picard et al., datasets).

In the Besnard et al., dataset, we find that the largest inter-RO distances in each human cell-line are 3.59 Mbp (IMR90), 3.71 Mbp (hESC), 3.71 Mbp (iPSC), and 4.29 Mbp (HeLa); and in Picard et al., dataset, the largest inter-RO distances are 5.65 Mbp (IMR90), 5.73 Mbp (HeLa), and 5.94 Mbp (K562). Indeed, the largest inter-RO distances in all the human cell lines appear to be bounded by approximately one half of the stalling distance N_s, which means the largest inter-RO distance in each human cell line contributes approximately 5% of the genome-wide error rate while in yeasts the contribution from the largest inter-RO distance to the over all genomic error rate is less than 1%. Given that errors are very likely in human genome, we can determine the range of inter-RO distances that contribute most to the genome wide probability of DFSs. We grouped the inter-RO distances into five different cohorts: very small (XS; <1 Kbp), small (S; 1-10 Kbp), medium (M; 10-100 Kbp), large (L; 100 Kbp-1 Mbp) and very large (XL; >1 Mbp). The frequency of inter-RO gaps in these cohorts is shown for IMR90 from the Besnard et al., and Picard et al., data in Figure 11a and 11b. The most common range of inter-RO gaps is small and medium respectively, the shift from 'small' to 'medium' being due to the coalescence of small gaps in the Picard et al. study. 'Large' and 'very large' gaps appear only at low frequency. Despite this, Figure 11c and 11d show that the cohort of 'large' gaps dominates as the source of error, which is due to the fact that the DFS probability increases non-linearly with the length of the inter-RO gap (Figure 9d). The error rate due to the small number of 'very large' gaps is significantly smaller, but non-negligible. Similar results are observed in all other human cell lines in both datasets as been presented in respective sections in Figures 12, 13 and 14.

4.3.4 Large gaps in human genome are distributed as a power law

For a more extensive analysis of the inter-RO gaps contributing most to error in human cell lines, we checked the error rates in the 'large (L)' gap cohort and gaps in its proximity. Very interestingly, in all cell lines, error rates in the vicinity of the 'large (L)' gap cohort shows a surprisingly statistically uniform distribution of error rate, and this suggests ROs are placed so as to minimize the damage from error by means of spreading out the error across size scales. In Figures 10e and 10f, the probability of DFS in each 10 kbp interval in the range 10 - 300 Kbp is shown for the Besnard et al., (Figure 11e) and Picard et al., (Figure 11f) datasets for primary IMR90 cells. These are the inter-RO gaps that contribute the most to the genome wide DFS probability. The maxima are relatively broad, particularly for the Besnard et al. dataset; the probability of DFS in each 10 Kbp size range is approximately constant at 0.030-0.035 across inter-RO distances spanning from 40 Kbp to 200 Kbp and for the Picard et al. dataset, the probability of DFS in similar size range is approximately constant at 0.040-0.050 across inter-RO distances spanning from 30 Kbp to 120 Kbp. In order to maintain this similitude in error rate despite the increasing size range, the frequency of the gaps in each size range must fall appropriately. To check this relationship between the frequency of inter-RO gaps close to the vicinity of a given gap size we derived the following formula given by Eq. ($B17$) in chapter 2:

$$M = \frac{\log(\theta)}{[\log(1+Nq) - Nq]} \qquad (R9)$$

where M is the frequency of inter-RO gaps, N is the gap size, q is the per nucleotide error rate given by '$\log(2)/N_s$' and θ is the probability of no error occurring from the inter-RO gaps counted by M which are all close to N in size and in chapter 2, Eq. (B18) we have shown that this provides the following relationship:

$$M \sim \frac{1}{N^2} \qquad (R10).$$

Thus, for inter-RO distances significantly smaller than the stalling distance N_s, we can infer that ROs are placed in such a way to give a power law, with a frequency of DFSs that decreases as the inverse square of the length of the inter-RO distances thereby spreading the probability of a DFS equally amongst all size classes.

The frequency of inter-RO gaps decreases to balance the effect of the larger gap size maintaining a similar error rate across the gap size classes. Figure 11g and 11h show that there is a remarkable concordance between the theoretical frequency distribution (in blue) with the frequency distribution in the data for IMR90 cell-line in both Besnard et al., and Picard et al., datasets (in red). There is also similar agreement with the theoretical distribution of the frequency of inter-RO gaps close to the 'large (L)' gap cohort in all the other cell-lines in both datasets as shown in respective figures in Figure 12, 13 and 14. These results can be interpreted in terms of "spreading the damage" as widely as possible in the inter-RO gap region of maximal DFS errors, as a power law is the most effective way to delocalize errors from any single cohort of inter-RO distances.

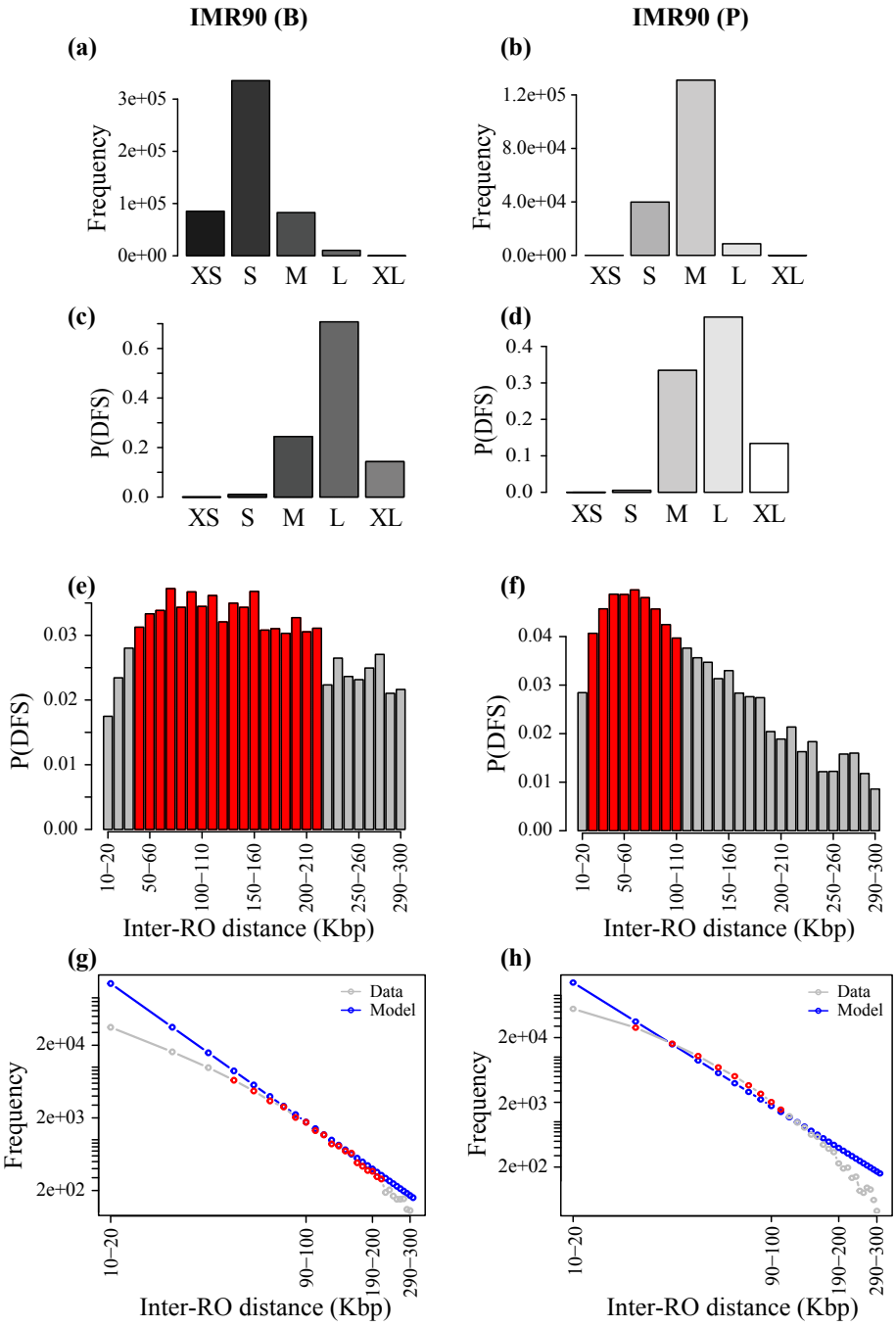


Figure 11: Data in the left and right columns is from the IMR90 human datasets 'B' (Besnard et al.,) and 'P' (Picard et al.,) respectively; a & b) Frequency of replicons in each cohort; defined according to the following size ranges, $<10^3$ bp = XS, 10^3–10^4 bp = S, 10^4–10^5 bp = M, 10^5–10^6 bp = L, $>10^6$ bp = XL. c & d) Probability of DFS in each cohort of the inter-RO distances; e & f) Higher resolution plot of probability of DFS at the transition from "medium (M)" to "large (L)" gap cohorts, contributing most towards the P(DFS); red bars show the bins with maximum P(DFS) in respective datasets and the contingent of bars with very similar P(DFS)s at the peak of the distribution of errors from each 10 Kbp cohort are marked as the maximum error contributing cohorts; g & h) Theoretical frequency distribution of replicons inferred from the plots e & f are presented in blue; grey shows the actual frequency distribution in those bins in the data and red highlights the red bins in e & f.

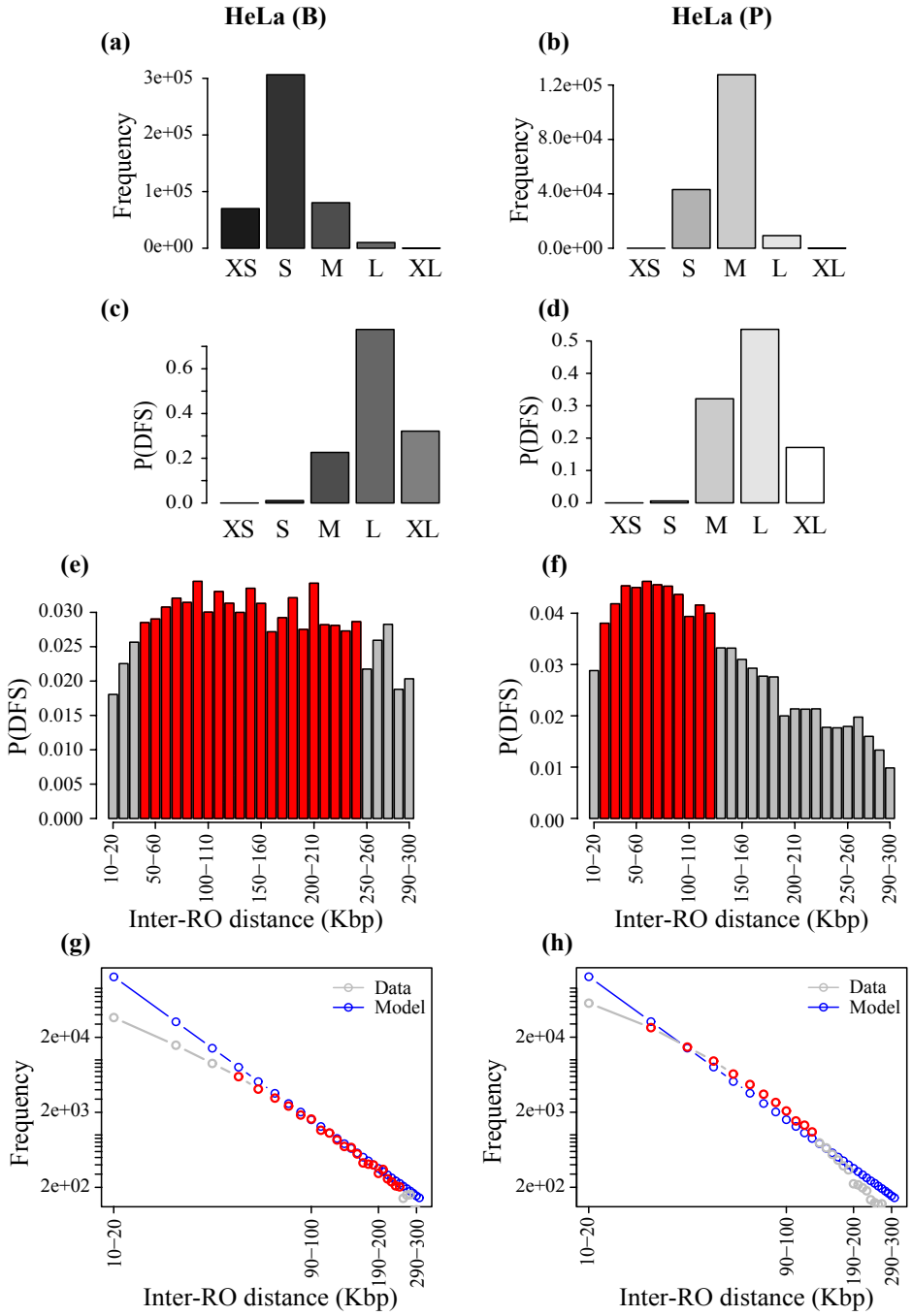


Figure 12: Data in the left and right column is from HeLa human dataset B and P respectively; a & b) Frequency of replicons in each cohort; defined according to the following size ranges, $<10^3$ bp = XS, 10^3–10^4 bp = S, 10^4–10^5 bp = M, 10^5–10^6 bp = L, $>10^6$ bp = XL; c & d) Probability of DFS in each cohort of the replicons. e & f) Higher resolution plot of probability of DFS at the transition from "medium (M)" to "large (L)" gap cohorts, contributing most towards the P(DFS); red bars show the bins with maximum P(DFS) in respective datasets and the contingent of bars with very similar P(DFS)s at the peak of the distribution of errors from each 10 Kbp cohort are marked as the maximum error contributing cohorts; g & h) Theoretical frequency distribution of replicons inferred from the plots e & f are presented in blue; grey shows the actual frequency distribution in those bins in the data and red highlights the red bins in e & f.

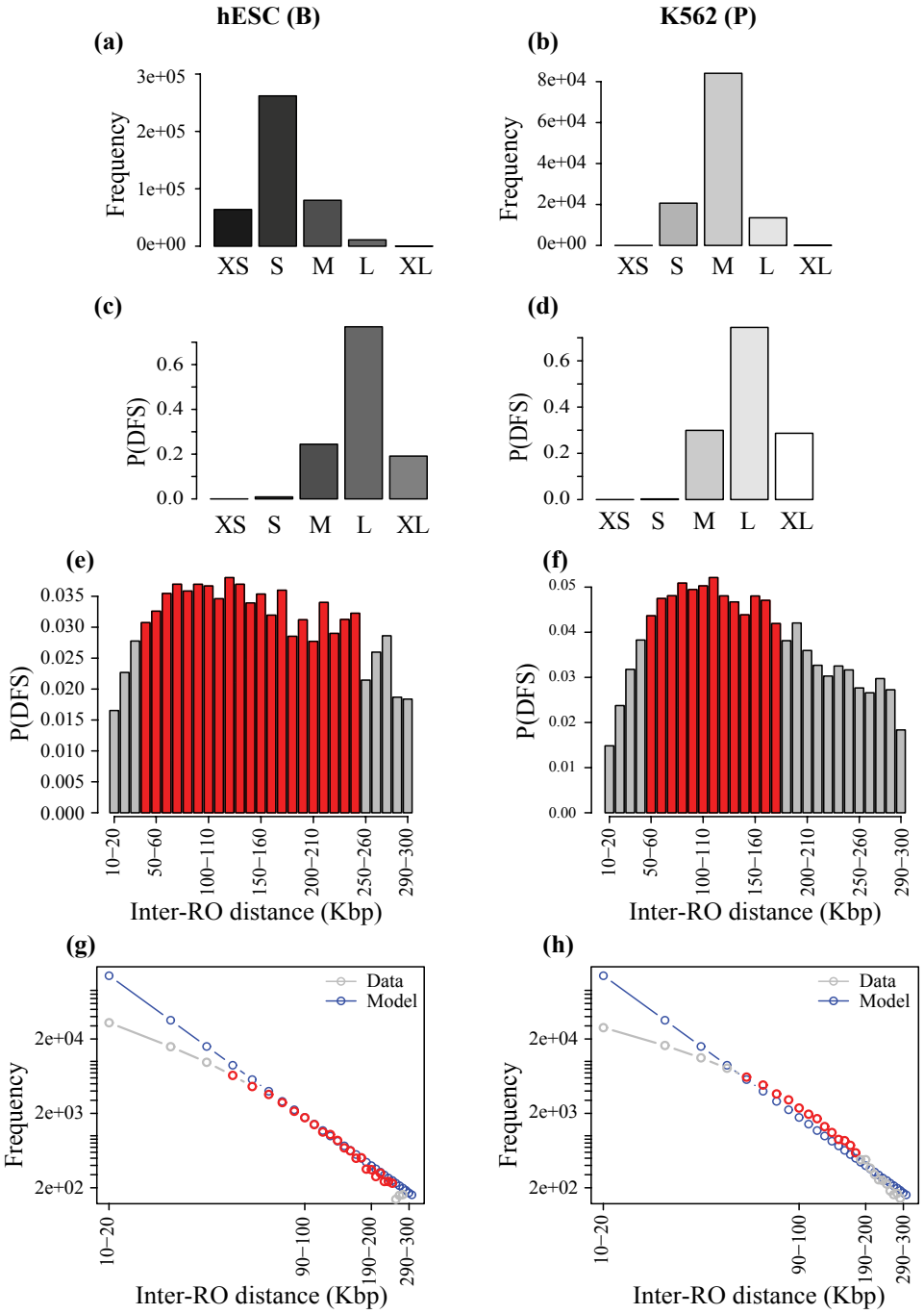


Figure 13: Data in the left and right column is from hESC and K562 in human dataset B and P respectively; a & b) Frequency of replicons in each cohort; defined according to the following size ranges, $<10^3$ bp = XS, 10^3–10^4 bp = S, 10^4–10^5 bp = M, 10^5–10^6 bp = L, $>10^6$ bp = XL; c & d) Probability of DFS in each cohort of the replicons. e & f) Higher resolution plot of probability of DFS at the transition from "medium (M)" to "large (L)" gap cohorts, contributing most towards the P(DFS); red bars show the bins with maximum P(DFS) in respective datasets and the contingent of bars with very similar P(DFS)s at the peak of the distribution of errors from each 10 Kbp cohort are marked as the maximum error contributing cohorts; g & h) Theoretical frequency distribution of replicons inferred from the plots e & f are presented in blue; grey shows the actual frequency distribution in those bins in the data and red highlights the red bins in e & f.

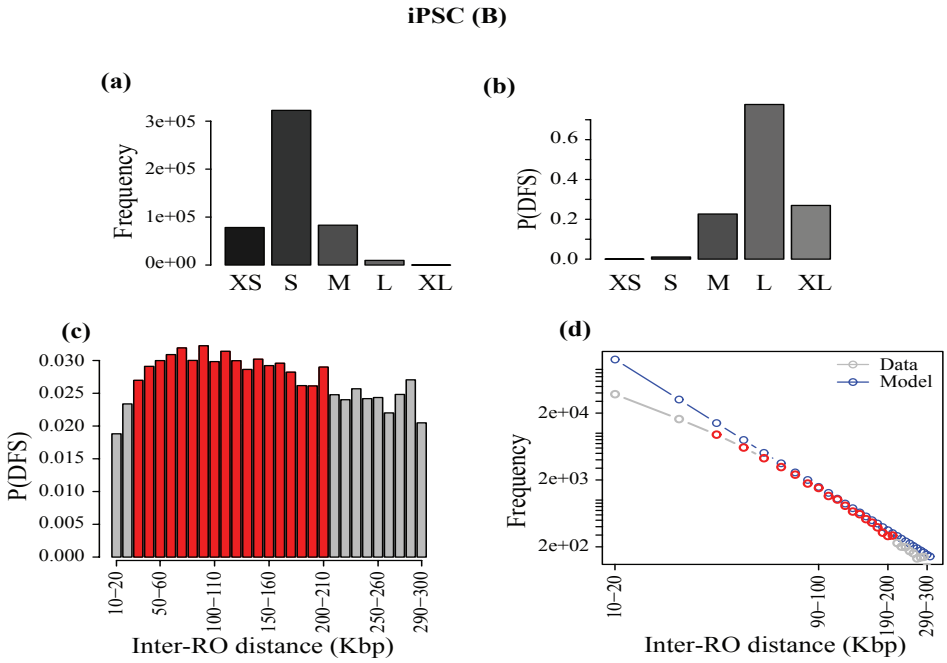


Figure 14: Data is from iPSC human dataset in B; a) Frequency of replicons in each cohort; defined according to the following size ranges, $<10^3$ bp = XS, 10^3–10^4 bp = S, 10^4–10^5 bp = M, 10^5–10^6 bp = L, $>10^6$ bp = XL; b) Probability of DFS in each cohort of the replicons. c) Higher resolution plot of probability of DFS at the transition from "medium (M)" to "large (L)" gap cohorts, contributing most towards the P(DFS); red bars show the bins with maximum P(DFS) in respective datasets and the contingent of bars with very similar P(DFS)s at the peak of the distribution of errors from each 10 Kbp cohort are marked as the maximum error contributing cohorts; d) Theoretical frequency distribution of replicons inferred from the plots e & f are presented in blue; grey shows the actual frequency distribution in those bins in the data and red highlights the red bins in c.

4.3.5 Replication errors are common but low in higher eukaryotes and are distributed as Poisson

We have shown in chapter 2 under '**Model B**', our theory predicts that the distribution of the number of DFSs in a given genome is Poisson-distributed to a very high degree of accuracy. We have applied our theory to the human cell lines datasets to test this prediction. As shown in Figure 15, for all cell lines, in both Besnard et al., and Picard et al., datasets, the distribution of DFSs is indeed Poisson-distributed, regardless of being primary or tumoural cell lines. Statistical analysis confirms that the computationally derived probability distribution of DFSs is statistically indistinguishable from the fitted Poisson distribution. Interestingly, we find a very low probability (<10%) of encountering more than three DFSs in the replication of the

entire diploid human DNA per cell cycle. Therefore, despite the high probability of the presence of DFSs (~80%), in ~90% of cells undergoing DNA replication the expected number of DFSs is predicted to be three or less, with one or two errors being the most likely occurrences. Indeed, we find that the parameter λ (i.e., the mean number of errors) that characterizes the distribution of DFSs ranges from 1.67 to 2.15 in Besnard et al., and from 1.21 to 2.05 in Picard et al., datasets. Given that DFSs in human cell lines are almost inevitable, it is surprising to find that their number is quite sharply constrained to be essentially one, two or three. This might indicate that the mechanism that deals with such errors has a very low capacity. DFSs are the primary cause of DNA double strand breaks during replication (Allen et al., 2011; Jones et al., 2014; Unno et al., 2013), and are likely to be major contributors for the development of cancer and other pathologies, such as ones associated with aging (Bohgaki et al., 2010; Li et al., 2008). The inevitability of DFSs in longer genomes requires the presence of cellular mechanisms, which are able to deal with such errors in an efficient manner.

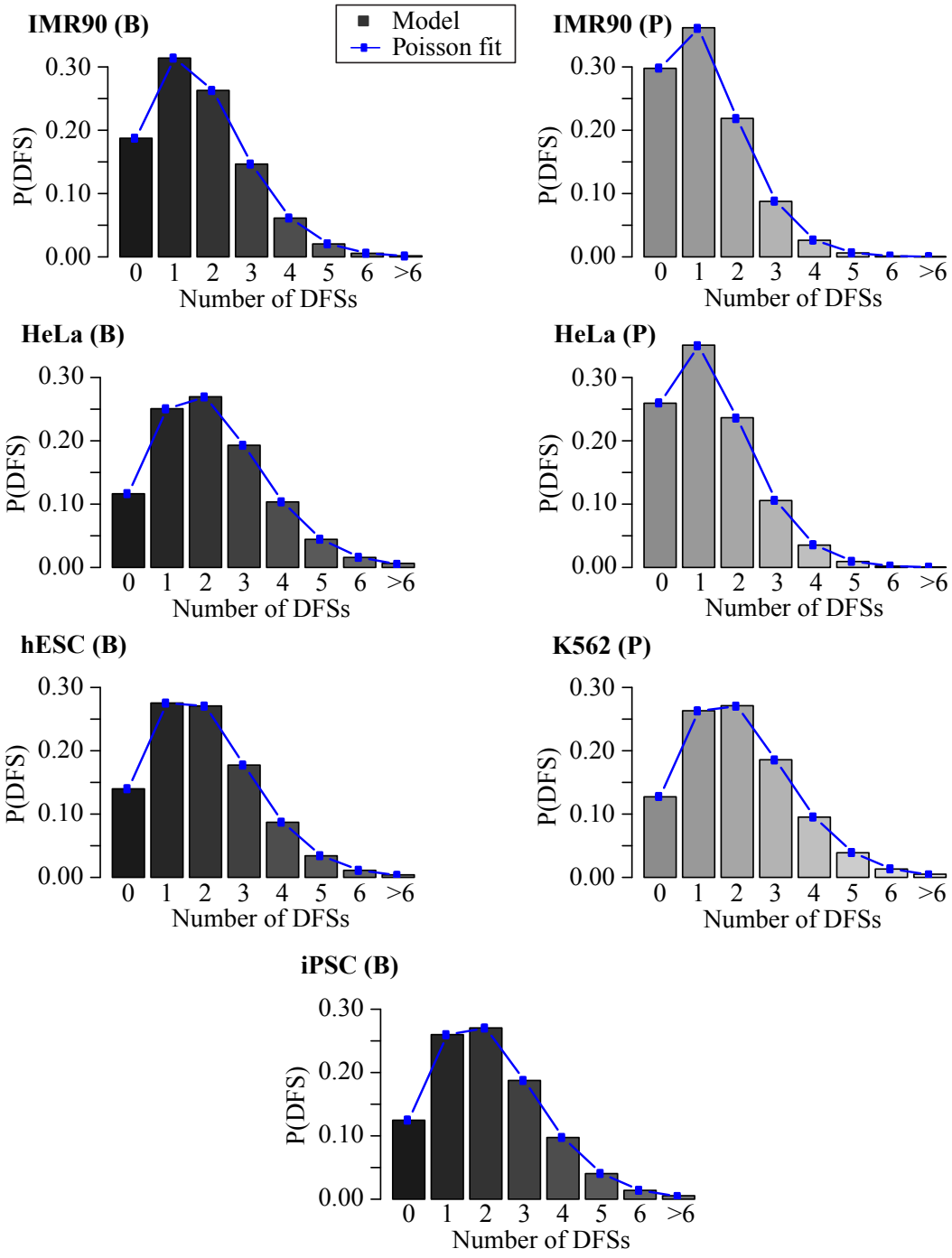


Figure 15: Theoretical prediction for the distribution of the number of DFSs based on the RO positions in each human cell-line datasets (using data from both 'B' and 'P'); also shown, as lines and dots, are best fits to a Poisson distribution.

There is very recent experimental evidence for one such post-replicative mechanism, involving the segregation of unreplicated DNA via UFBs and its protection by 53BP1 before being resolved in the next S phase (Moreno et al., 2016). In this mechanism

there are two different “biomarkers” for double strand breaks which would arise from DFS errors: these are the presence of 53BP1 nuclear bodies in the G1 phase of the subsequent cell cycle, and the presence of ultrafine anaphase-bridges (UFBs) during mitosis. Our theory suggests that the number of both 53BP1 nuclear bodies and UFBs are distributed as a Poisson with a value of λ between one, two or three. If, as suggested in (Moreno et al., 2016), the defects induced by DFSs can be resolved in the following cell cycle by segregating unreplicated DNA to daughter cells, DNA strand breaks could be generated at each DFS. Because the number of illegitimate ways that double strand breaks could be correctly rejoined increases as the factorial of the number of breaks, this might constrain the number of tolerated DFSs to about 3 or less.

Our experimentalist collaborator Professor Julian Blow’s lab has done the experiment to measure the frequency of 53BP1 in IMR90 cells and both 53BP1 and UFBs in U2-OS cells, and measured the frequency of their occurrence during the cell cycle at a single cell level (Moreno et al., 2016). We were privileged to have the data from Professor Blow’s group in order to check our theoretical predictions. In agreement with our predictions, the experimental distributions of both 53BP1 nuclear bodies and UFBs fit to a Poisson distribution (Figures 16a, 16b and 16c). Statistical analyses indicate that both a naïve fitting using the mean of the data and a more advanced approach that accounts for potential errors introduced by the experimental procedure of the immunofluorescence experiments (Figures 16a, 16b and 16c) produce distributions which are not statistically different from Poisson distributions for both 53BP1 nuclear bodies (P values between 0.61 and 1 for both IMR90 and U2-OS cells) and UFBs (P values between 0.53 and 1 for U2-OS cells). Additionally, the fitted λ

values, 0.52 (naïve) and 0.54 (filtered) in IMR90 and 1.64 (naïve) and 1.89 (filtered) in U2-OS cells for 53BP1 nuclear bodies, and 1.27 (naïve) and 1.19 (filtered) for UFBs, are in line with the expectation of a limited number of DFSs. Moreno et al (2016) show that the number of 53BP1 nuclear bodies and UFBs follows a Poisson distribution in the HeLa cell line with λ values of 0.94 (naïve) and 1.12 (filtered) for 53BP1 and 1.43 (naïve) and 1.19 (filtered) for UFBs (Moreno et al., 2016). Taken together, these results provide good agreement of our theory with the available data and reinforce the connection between 53BP1 nuclear bodies and UFBs to DFSs. The analysis of UFBs in unperturbed IMR90 cells was not possible due to experimental difficulties related to the fact that this cell line is not immortalized. As a more direct quantitative analysis, we compared the λ values obtained by direct calculation from the RO distribution of different human cell lines and the experimental λ values estimated from the distribution of 53BP1 and UFBs from the data shared by Professor Blow's lab. Note that comprehensive RO distribution data are not available for the cell line used for the UFB experiments (U2-OS) and diversity has been observed in RO-distribution across different cell lines (Besnard et al., 2012). Moreover, both 53BP1 and UFBs are likely to provide only an approximation of the number of DFSs as they appear also in the presence of non-DFS associated double strand breaks. Despite these limitations, a comparison of the λ values indicates that experimental measures are in excellent agreement with theoretical prediction (Figure 16d).

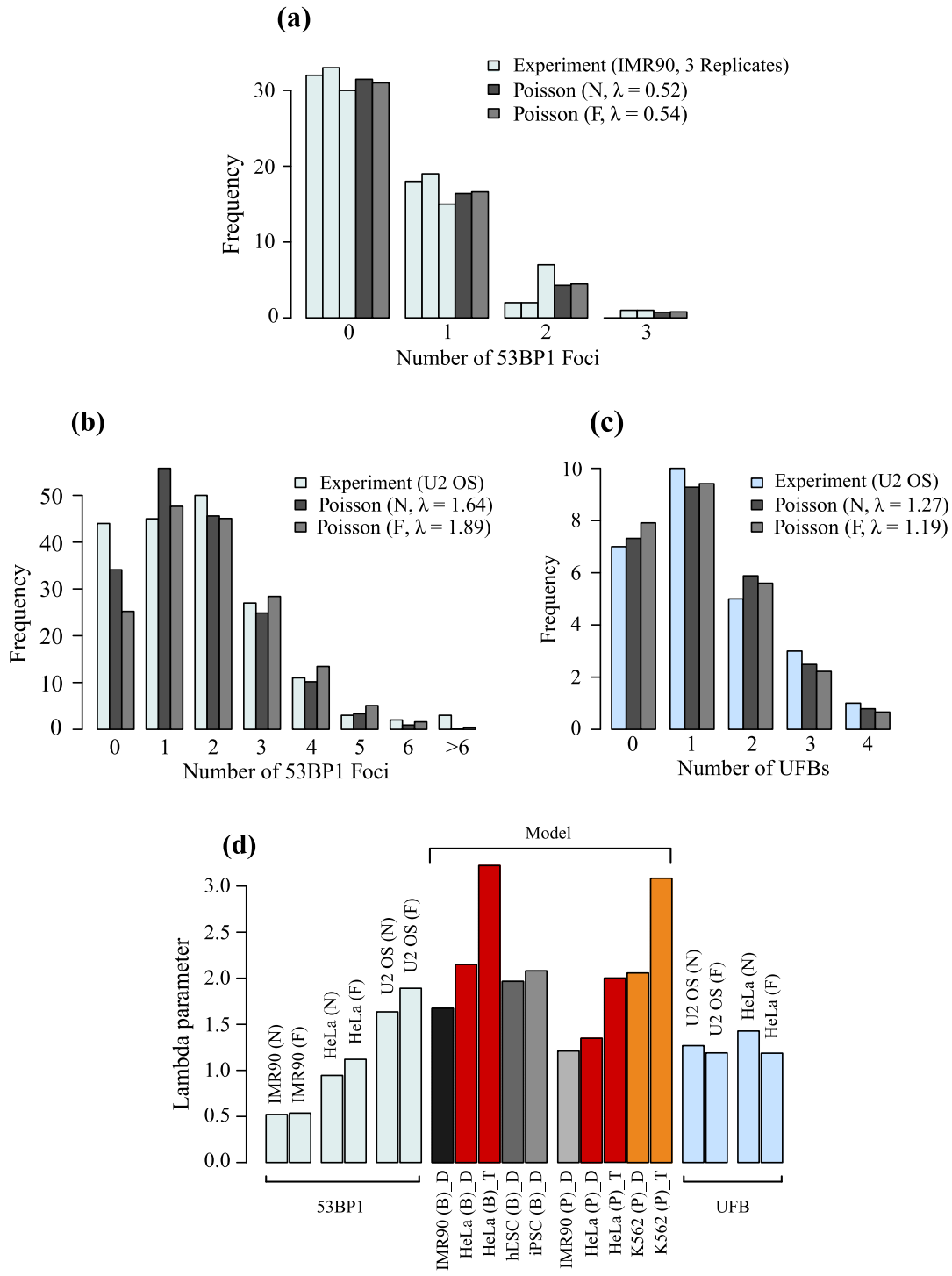


Figure 16: a & b) Experimental distribution of 53BP1 nuclear bodies in the IMR90 cell-line (3 replicates) and in the U2-OS cell-line (1 replicate) fitted with a naïve Poisson (i.e. taking the mean of the data as λ) (gray) and a filtered Poisson (i.e. ignoring the frequncies of zero counts to account for potential error from immunofluorescence staining) (lightgray). c) Experimental distribution of UFBs in the U2-OS cell-line fitted with a naïve Poisson (gray) and a filtered Poisson (lightgray). d) Values of the Possion parameter λ obtained from experimental fits of 53BP1 in IMR90, HeLa, U2-OS and UFBs in U2-OS, HeLa are compared with theoretical

values obtained from different cell lines in Figure 15. Under the theoretical Model values: D denotes diploid and T denotes triploid as HeLa and K562 could be hypotriploid in cell cultures. Note: the experiments were carried out by Alberto Moreno in Professor Blow's lab.

Additional comparisons with the λ values obtained from HeLa (Moreno et al., 2016) reinforce our predictions to be experimentally valid (Figure 16d). Interestingly, the range of variation observed in the experimental value of λ is matched by the range of variation of our model predictions, suggesting that our methodology is correctly capturing experimental variations. All together, there is a very good agreement in the numbers and statistical distribution of experimental measurements of both 53BP1 and UFBs with the predictions of Poisson statistics from our theory, supporting the validity of our conclusions. Plausibly, this mechanism involving 53BP1 and UFBs for resolving DFSs has limited capacity. Thus, our theory, in light of the experimental data, shows a contingent trade off between inevitability of DFS occurrence and the difficulty of its resolution.

In both IMR90 and HeLa cells the experimentally derived λ obtained from 53BP1 nuclear bodies data is approximately half of the theoretical estimate obtained from the RO mapping data. This is also true for UFBs in HeLa cells. In order to emphasize the implication of this observed small difference, even though the data and theoretical predictions are very close, we wanted to check the relationship with the value of λ to the density of licensed ROs on the genome. We can check this by considering the formula for no error genomewide in Model C in the context of Poisson parameter λ as shown in Eq. (*R*8). In Model C as given by Eq. (*C*6),

$$P(\text{error}) = 1 - \exp\left[-U\,\frac{N_g(2-\rho)}{\rho}\right]$$

just to remind the reader from chapter 2, here U is the replication constant that was introduced in chapter 2 Model C and we will discuss this in detail in chapter 5. N_g is the genome length and ρ is the density of ROs on the genome. Now,

$$P(\text{zero error}) = \exp\left[-U\,\frac{N_g(2-\rho)}{\rho}\right].$$

In a Poisson distribution, $P(\text{zero error}) = \exp[-\lambda]$, and we have shown already that the replication errors in our consideration are Poisson distributed. Hence, we write

$$\lambda = U\,\frac{N_g(2-\rho)}{\rho}.$$

and this gives us a direct relationship between mean error λ and density of licensed ROs ρ. Now, for $\lambda' = \frac{1}{2}\lambda$, we have

$$\lambda' = \frac{1}{2}\lambda\,,$$

or,

$$U\,\frac{N_g(2-\rho')}{\rho'} = \frac{1}{2}U\,\frac{N_g(2-\rho)}{\rho}$$

$$\frac{(2-\rho')}{\rho'} = \frac{(2-\rho)}{2\rho}$$

$$\frac{2}{\rho'} - 1 = \frac{(2-\rho)}{2\rho}$$

$$\frac{2}{\rho'} = \frac{2+\rho}{2\rho}$$

or,

$$\rho' = \frac{4\rho}{2+\rho}$$

As long as, ρ is small it is straightforward to write

$$\rho' = 2\rho\,.$$

Thus, doubling the density of ROs halves the value of λ. Hence the factor of two differences in the experimental and theoretical values of λ could indicate that around half of the genomic ROs are missing in the current datasets (e.g. due to difficulties in detecting ROs that fire very rarely or ROs positioned in repetitive regions of the DNA). This line of reasoning is also consistent with a potential issue with the largest measured inter-RO gap being approximately 4 Mbp; the issue being that the replication time for such a gap would be significantly longer than typical S-phase (ca. 8 hours) (Cooper, 2000). If the true RO density is twice that measured, one can show that the largest gap would be halved, giving a value of 2 Mbp which is in line with the estimate of 2 Mbp for the longest stretch of DNA that could be replicated in the duration of S-phase (assuming a fork speed of approximately 2 Kbp per minute (Méchali, 2010), and remembering that a large replicon will be replicated almost symmetrically by forks travelling from either end). HeLa and K562 cell lines are highly aneuploid in culture and to account possible aneuploidy we calculated the theoretical λ values for triploid conditions of these two cell lines. The model predicted λ values in triploid HeLa and K562 cell lines around 3 and less in both datasets are within the expectation of low λ (between one and three). Apparent three-fold difference between experimental and theoretical λ in HeLa triploid scenario (Figure 16d), reemphasizes the issue of missing ROs. Mathematically it is straightforward just as before in the case of doubling ρ minimizes λ by one-half; tripled RO density ρ makes the mean error λ one-third. Hence, the model predictions show strong consistency in the face of missing ROs as well as large scale aneuploidy. It is worth stressing that our central equation for λ, the mean number of DFSs, contains very large numerical values, i.e. N_g and N_s, as well as thousands of inter-RO distances. Therefore, in principle, the formula could have produced values for λ of almost

arbitrary magnitude, either much less than or much greater than unity. It is striking that our theoretical predictions from the central equation yield values for λ close to unity and in such strong agreement with experimental data.

4.3.6 Estimation and effect of variation of stalling distance

In applying our theory to the RO position data for various human cell lines, we can vary the numerical value of the median stalling distance N_s and measure the effect on the expected number of DFSs. This allows us to check the extent to which our conclusions are robust to the variation of the only parameter in our analysis for which we do not have strong experimental data. Both theoretical and biological estimates indicate that N_s is approximately 10 Mbp (Maya-Mendoza et al., 2007; Newman et al., 2013). However, a precise estimate of this value is difficult to determine *in vivo*. The stalling distance is inversely proportional to the very small probability of an irreversible stalling event per nucleotide replicated, which because of the conservation of the basic replication machinery is likely to be relatively conserved across eukaryotes.

First, we analyzed the overall probability of DFSs occurring as N_s is varied. In all the human cell lines considered we observe a characteristic transition around 5 Mbp: below this value the probability of observing DFSs saturates at one (Figure 17a). Therefore, DFSs are inevitable for smaller values of N_s as one might expect. Importantly, our analysis indicates diminishing returns when N_s is increased to much larger values: even for N_s around 30 Mbp, error rates are sufficiently high (1 in 5 cells would experience a DFS during S phase) that additional DFS repair mechanisms are still required.

Therefore, in higher eukaryotes with large genomes the pressure to maintain genome stability is most easily resolved by additional safeguard mechanisms to deal with consequences of DFSs, rather than by stabilizing the replication machinery to give such a large N_s that DFSs can be avoided with the regular RO distribution found in eukaryotes with smaller genomes. Our analysis stresses the inevitability of DFS errors during replication of the human genome and calls for a shift in our approach with respect to how the problem has been viewed in the past. On varying the median stalling distance in human cells, the probability of exactly one DFS genome-wide reaches a maximum between 10 and 15 Mbp, depending on the particular cell line and dataset used (Figures 17b and 17c). Furthermore, on varying the stalling distance, we find that the probability of exactly two or exactly three DFSs occurring also have peaks in the range 6-10 Mbp, again depending on the cell line and the dataset used (Figures 17b and 17c). To probe the likelihood of small number of errors occurring, we plotted the probability of observing *one, two or three* DFSs as stalling distance was varied (Figures 17d and 17e). These results show a very pronounced maximum for N_s around 10 Mbp in the Besnard et al., dataset, and around 8 Mbp in the Picard et al., dataset. In summary, our analysis of the available RO distribution in a variety of human cell lines and in different datasets indicate that only for N_s in the vicinity of 10 Mbp the number of DFSs is constrained between zero and three.

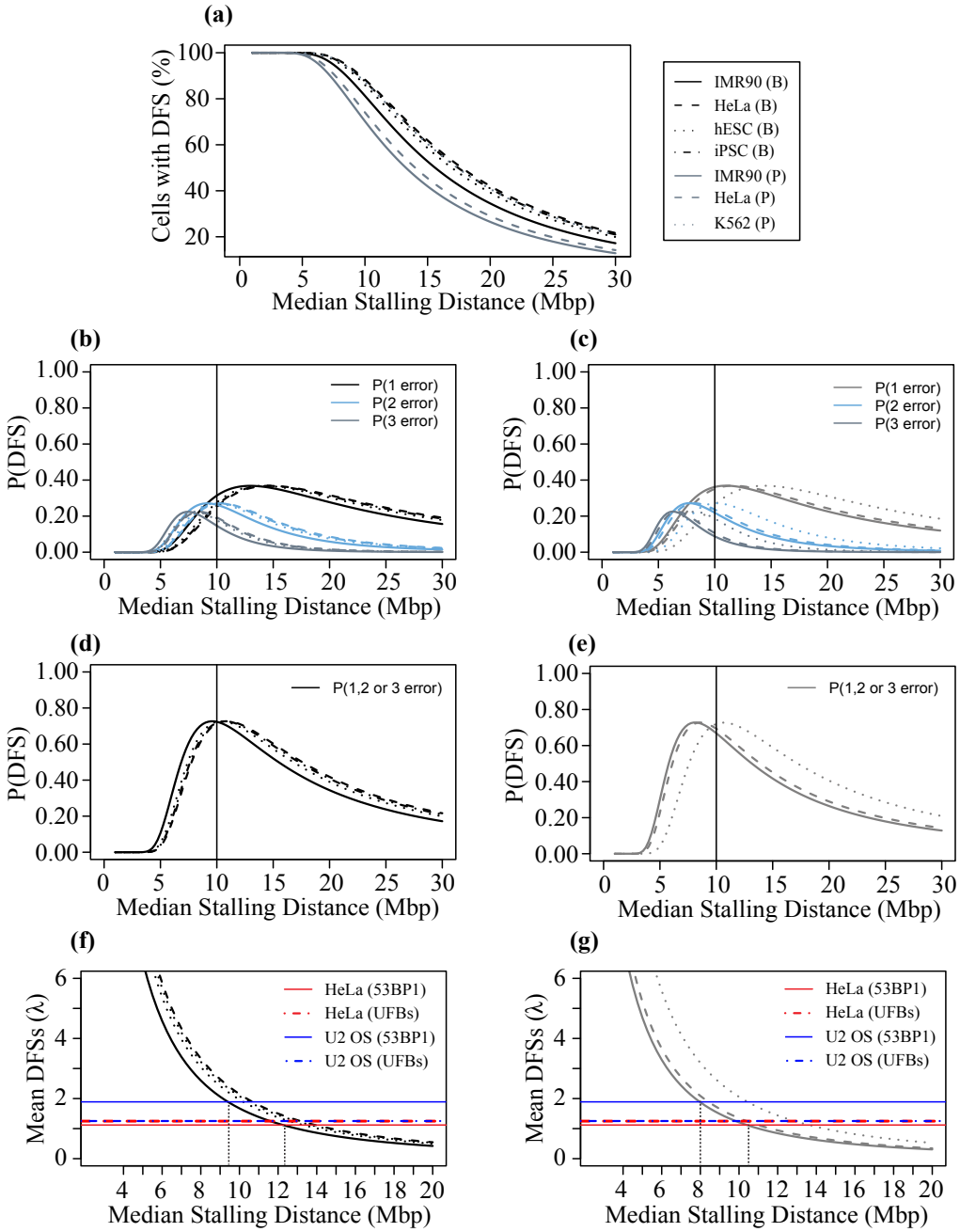


Figure 17: a) Based on the RO distributions in the various human datasets, theoretical predictions of the percentage of cells with DFSs is plotted as a function of the parameter N_s (median stalling distance); the percentage is essentially 100% when N_s < 5 Mbp and this percentage is still non-trivially high even when N_s > 20 Mbp. b & c) Theoretical predictions of the probability of one, two and three DFSs is shown as a function of N_s. d & e) Theoretical predictions of the probability of one, two, or three DFSs is shown as a function of N_s. f & g) Expected numbers of DFSs in different cell-lines are plotted against N_s; in blue and red are the experimentally obtained expected number of 53BP1 nuclear bodies and UFBs in U2-OS and HeLa cell-lines respectively. Crossing points of the blue and red lines over the curves provide an independent estimate for the plausible range of N_s (vertical lines) by directly comparing experimental data with theoretical predictions.

Finally, we can measure the average number of DFSs when N_s is varied. This number is equal to the λ parameter of a Poisson distribution, and therefore allows a direct comparison to the experimental measures that we have discussed in the previous section. As expected, the average number of DFSs decreases from a large value as N_s is increased (Figures 17f and 17g). As explained in the previous section, fitting the Poisson distribution to 53BP1 and UFB experimental data (obtained from Professor Julian Blow's lab) gives values of λ between 1.12 and 1.89 (the values are shown in Figures 17f and 17g as blue and red lines). The intersection of the decaying curve with these two lines provides another independent estimate of the stalling distance N_s, which we find to be between 8 and 13 Mbp depending on the cell line and dataset used. Our analysis of the statistics of DFSs in human cell data on varying the stalling distance therefore provides very strong evidence for the robustness of this parameter with a value of approximately 10 ± 2 Mbp, consistent with previous estimates from our analysis of yeast RO distributions, and direct experimental estimates (Maya-Mendoza et al., 2007; Newman et al., 2013).

4.3.7 Effect of varying the number of licensed ROs

Interestingly, amongst the cell types we analyzed, there was no major difference in the mean inter-RO distance (Figure 9b). Figure 18 shows how decreasing mean inter-RO distance would reduce the probability of DFSs in a *generic* organism. The black, light-blue, and blue lines illustrate the mean inter-RO distance to achieve a fixed probability of DFSs under the optimal situation of equally spaced ROs. All the datasets analyzed in the article have a mean inter-RO distance ranging between 10 and 100 Kbp (shaded pink in Figure 18). Because of the relatively small genome sizes of yeasts, so long as ROs are evenly spaced this mean inter-RO distance can achieve a

tolerable DFS probability of ~0.1%, similar to the chromosome missegregation rate as shown in chapter 2 (Newman et al., 2013). In order to maintain a low probability of DFSs as in yeasts, longer genomes would require a much lower mean inter-RO distance or in other words, much higher density of ROs on the genome. Since the MCM2-7 double hexamer that licenses an RO has a footprint of ~60 bp (Evrin et al., 2009; Remus et al., 2009) this provides an absolute limit to the possible inter-RO distance (dashed line in Figure 18). It is just about possible for organisms with ~6,000 Mbp genomes to achieve yeast-like DFS probabilities, but the genome would have to be almost completely packed with MCM2-7, which might leave the genome unable to perform its major function of providing the template for transcription. Since this is an implausible saturation for normal cells, additional post-replicative mechanisms must be in place to deal with the inevitable DFSs. For this reason, regularity in RO distribution is not an effective safeguard against DFSs in organisms with larger genomes.

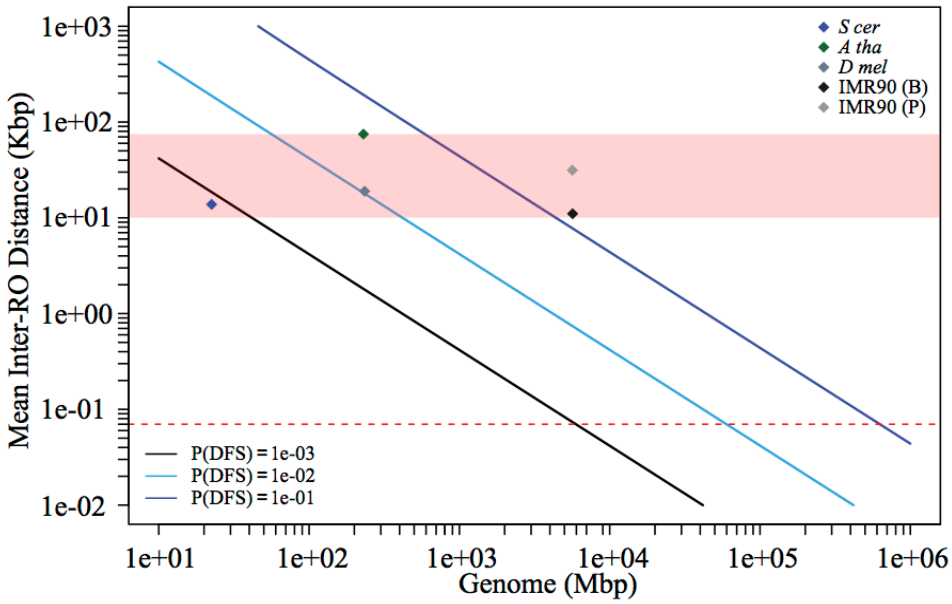


Figure 18: Highlighting the issues faced to maintain small DFS error rates for genomes of increasing length: Theoretical prediction of the average replicon length as a function of increasing genome length, to maintain a fixed probability of DFS, for three different values of this probability; diamonds show the positions of yeast, *Arabidopsis, Drosophila* and human respectively, obtained from the datasets of RO positions. The pink shadow highlights the biologically relevant range for mean replicon lengths as per all eukaryotic datasets available. The dashed red line marks the footprint for the MCM2-7 double hexamer, below which any replicon length is biologically unrealistic.

4.4 Discussion

Complete and faithful DNA replication is an absolute necessity to preserve the genetic content of organisms and impaired DNA replication may cause severe consequences. As ROs can not be licensed after replication initiation, to avoid double replication which is other side of the coin for incomplete replication and has severe consequences too, ROs must be licensed in appropriate location that minimizes the replication error. In this chapter we have shown how understanding the principles that govern distribution of ROs provides new quantitative insights into the way that different organisms maintain genetic integrity. By using a probability theory approach, based on a one-parameter model with simple yet plausible assumptions, we

have developed a set of measures and predictions that further this understanding. Remarkable agreement of our theoretical predictions with experimental data strongly supports the validity of our model assumptions. Additionally, it allows us to explore the rich system-level diversity of features and constraints associated with DNA replication in eukaryotes spanning simple yeasts to complex higher eukaryotes. The extent of sophistication in phenotypic complexity of organisms is generally associated with genome length and complex metazoans have much larger genomes compared to yeast: the diploid human genome is approximately 600 times larger than the haploid yeast genome. Despite this large difference in genome length, the replication machinery is essentially conserved throughout the eukaryotes (Sclafani and Holzen, 2007). Over the past few decades, much research has been done in understanding the molecular mechanisms involved in eukaryotic DNA replication and the associated damage-repair mechanisms. However, less is understood about the system-level structures, processes and physical features that ensures replication fidelity across the different scales of eukaryotic complexity, mirrored by genome lengths spanning over three orders of magnitude across yeast to human. In our model we consider the double fork stall (DFSs), one of the major threat to complete genome duplication in eukaryotes and check the implication of this error in different organisms with different genome lengths and how the replication fidelity is maintained in the face of such error events.

The 'central equation' in '**Model B**' shows that there is a hierarchy of contributions to the probability of DFS errors, with genome length being the most significant factor, followed by RO number and then RO distribution. The observed difference in these factors among the organisms considered effectively creates different classes of probabilities of DFS errors ($\sim 10^{-3}$, $\sim 10^{-2}$, and ~ 1) for the respective classes of

organisms based on their genome lengths (~20 Mbp, ~250 Mbp and ~6 Gbp). Interestingly, amongst the organisms we analyzed, there was no major difference in the density of ROs i.e. mean inter-RO distances. One possible explanation for this is that in order to make a significant effect on reducing DFSs, as reducing mean inter-RO distance by increasing the RO density over the genome would reduce the error rate, the RO density in organisms with genomes of ~250 Mbp or more would lead to excessive clashes with the transcriptional machinery. The third component of our equation – the nature of the RO distribution can be characterized by the measure for the degree of uniformity of inter-RO distances in the genomes, i.e. coefficient of variation in RO distribution, *R* also reflects these classes (with values <1, ~1 and >1 respectively) among the organisms, indicating that as the probability of DFSs approaches 1 in larger genomes, the pressure towards a regular RO distribution is progressively lost.

Due to the pathological and cell biological consequences associated with DFSs, inevitability of DFSs in longer genomes requires the presence of cellular mechanisms that are able to deal with such errors in an efficient manner. (Moreno et al., 2016) has shown evidence for such mechanisms involving 53BP1 nuclear bodies and UFBs as biological markers. We have demonstrated very good agreement in the numbers and statistical distribution of experimental measurements of both 53BP1 and UFBs with the predictions of Poisson statistics from our theory, supporting the validity of our conclusions. Analysis of the data available for human cell lines within our theoretical framework shows that RO density and distribution constrain the number of DFSs per cell cycle to three or less for nearly all cells and interestingly the mechanism suggested by (Moreno et al., 2016) has the capacity to deal with these limited number of DFS errors. Therefore, our theory, in light of the experimental data available,

shows a contingent trade off between inevitability of DFS occurrence and the difficulty of its resolution.

Another important requirement for the containment of replicative errors in larger genomes is an upper limit in the length of largest inter-RO distance. Longer inter-RO distances correspond to a higher probability of DFSs (Figure 9d). Our theory indicates that the largest tolerable inter-RO distance in human cell lines are bounded by ~$0.5\times N_s$, and interestingly the largest inter-RO gap found in experimental datasets are around $0.3\times N_s$. We have also analyzed human cell line data by varying N_s and we have shown that the probability of observing a number of DFSs equal to one, two or three is maximized for N_s in the region of 10 Mbp. This value for N_s is in excellent agreement with previous experimental and theoretical estimates in human cell lines and yeasts (Maya-Mendoza et al., 2007; Newman et al., 2013). Due to the universality of replication machinery across eukaryotes and the necessity of error containment in larger genomes, we suggest this N_s value (~10 Mbp) to be robust and universal in eukaryotes. A further signature of the containment of inevitable errors in longer genomes is the distribution of the risk among inter-RO gaps of different sizes: a relatively narrow range of inter-RO gaps (of size ~40 to ~200 kbp) are the highest contributors to overall DFSs in human cell lines and the different size ranges for inter-RO gaps within these cohort maintains an equal error rate by decreasing the frequency of inter-RO gaps as the inverse square of the increasing gap size.

As a concluding remark, it is worth stressing that some organisms, particularly plants, have very large genomes sizes even far beyond the genome length in human cell lines, with N_g as large as ~100 Gbp (Francis et al., 2008). Application of our model in these organisms would suggest that the number of DFSs becomes much larger than three,

and actually in the region of ten or more. It has been observed that the cell cycle length in plants undergoes a sharp increase as the length of the genome exceeds about 25 Gb (Francis et al., 2008), potentially reflecting the significantly greater burden of DFS detection and correction in these organisms and also the variable ploidity in such plants might be a result of improper resolution of these errors and its consequences. We currently do not have genome-wide RO distribution data for these organisms to test this idea, but this would provide further opportunities for gaining new understanding of the system-level strategies that eukaryotes employ to minimize replication errors.

Chapter 5

Universal replication constant in eukaryotes

5.1 Brief introduction

Despite the tremendous diversity and variation across different scales of life, something is absolutely common among all the living cells – the DNA molecule and its semiconservative replication. Every living organism on the face of the earth reproduces via transmission of genetic material stored in DNA. Hence, for species continuation, integrity of genetic information must be confirmed by maintaining sustained fidelity of DNA replication mechanism (Alberts et al., 2002; Bebenek, 2008). Eukaryotic DNA is packaged in a hierarchical chromatin structure that is crucial for nuclear activities like gene transcription, DNA replication, error-correction. Primary to this complex organisation is mostly conserved ~200 bp long nucleosomes (octamer histone core wrapped with 147 bp of DNA) connected to each other by linker DNA (average length ~60 bp) (Szerlong and Hansen, 2011). Thousands of ROs are licensed prior to the initiation of replication process in S phase of the eukaryotic cell cycles and nucleosome organisation is functionally related to origin activity (Alver et al., 2014; Deniz et al., 2016; Eaton et al., 2010) and at the same time provides a minimum boundary in inter-RO gaps as ROs cannot be loaded in the DNA wrapped around the nucleosome cores. Origin recognition complex (ORC) and origin licensing factor (MCM2-7 double hexamer) both have an effective footprint of ~50 bp and ~60 bp respectively which can be loaded in the inter-nucleosome linkers (Evrin et al., 2009; Remus et al., 2009; Speck et al., 2005). We discussed in chapter 4 that the higher eukaryotic genomes can effectively minimize

the replication error probability by reducing the inter-RO distances but together, the wrapped DNA around the nucleosome core and the footprint of RO licensing factors provide the shortest possible inter-RO distance which is comparable to the total nucleosome length of ~200 bp. In S phase, origins fire stochastically where each origin give rise to bidirectional pair of RFs that drives the polymerase through DNA. The distance that an individual fork can travel in absence of any replicative stress i.e. the fork stalling distance is a crucial parameter that puts boundary to the permissible maximum inter-RO separations as the origin spacing has to be always smaller than fork stalling distance in order to confirm complete replication faithfully. The experimental and theoretical estimates for the fork stalling distance is ~12 Mb and and this distance is the reciprocal of the per nucleotide stalling probability, which is ~5.8×10^{-8}. The conserved nature of replication machinery across eukaryotes suggests this per nucleotide fork-stalling probability on average is conserved too (Al Mamum et al., 2016; Leman and Noguchi, 2013; Maya-Mendoza et al., 2007; Newman et al., 2013).

In **'Model C'**, we used the nucleosome scale minimum possible distance between adjacent ROs to quantify the probability of DFS errors in a given genome and two biologically conserved factors i.e. nucleosome length and fork-stall probability per nucleotide together forms a mathematical constant in this model. We propose a universal replication constant based on these two conserved molecular biological measurable i.e. nucleosome length and fork stalling probability for all eukaryotes. We use a probabilistic model to relate this universal constant to the genome lengths and replication error in both embryos and non-embryos and we show various organisms from different phyla confirm to the proposed universal replication constant.

5.2 Results

5.2.1 The 'universal replication constant' in eukaryotes

It has been shown that yeasts can effectively minimize replication errors arising from fork failures by means of an equally spaced origin distribution. While organisms with larger genomes need to minimize the origin spacing to an extent of single nucleosome in order to maintain such lower rate of error in absence of complex error repair mechanisms (Newman et al., 2013). Strictly time bound early cell divisions in transcriptionally quiescent embryos do not permit accessibility to such mechanisms and very high origin density has been observed during early embryogenesis (Goldar et al., 2008; Hutchins et al., 2016). In situation with very high abundance of ROs in the genome, the lowest minimum of inter-origin separation cannot be less than the nucleosome size as origins can be loaded only outside the wrapped histone core and average linker space is indeed structurally compatible to the origin molecular footprints for such saturation. Therefore, considering the individual nucleosome length as the minimum possible inter-origin separation (mid-point to mid-point), we construct a model to quantify the replication error originating from fork-failures:

$$P(\text{error}) = 1 - \exp\left[-U\,\frac{N_g(2-\rho)}{\rho}\right] \qquad (R9).$$

Where N_g is the genome length, ρ is the probability for a nucleosome to have a licensed origin (or in other way ρ represents the density of ROs in the given genome) and the universal replication constant U is given by

$$U = \frac{1}{2}q^2 N_{nuc}$$

Here N_{nuc} represents the nucleosome length and q is the per nucleotide fork stalling probability. Since both N_{nuc} and q are measurable quantities associated with fundamental molecular biology, and are statistically conserved across all eukaryotes,

we can use their measured values to obtain U once and for all, thus giving U the status of a constant maintained universally across all eukaryotes irrespective of their genomic and architectural complexity. Using the known values for both N_{nuc} and q, which are ~200 bp and ~5.8 x 10^{-8} respectively, we have the universal constant $U = 3.33 \times 10^{-13}$.

In order to test this practically we rearrange Eq. (1) in the following form:

$$U_{calculated} = -\frac{\rho}{(2 - \rho)} \frac{\log[1 - P(\text{error})]}{N_g} \qquad (R10).$$

In transcriptionally active non-early-embryo cells or in unicellular eukaryotes where whole genome saturation with origins is not feasible due to transcriptional constrains, ρ is smaller than 1 and $P(\text{error})$ can be large depending on the genome length of the organism. Hence, in such cells Eq. ($R10$) is applied. We use experimentally obtained replication failure rates in different organisms in place of $P(\text{error})$ along with the measured ρ from genome-wide RO mapping data and the calculated $U_{calculated}$ values are in excellent agreement with the constant U (Table 1).

Table 1: $U_{calculated}$ values in non-embryonic cells are in conformity to the proposed constant U.

	N_g (Mbp)	ρ	$P(\text{error})$	$U_{calculated}$
Yeast	22.45	~0.015(Siow et al., 2012)	~0.00064[a] + $(3.3 \pm 0.3)\times10^{-4}$ [b] (Newman et al., 2013; Zhu et al., 2014)	~3.33×10^{-13}
Human	~6000	~0.006(Picard et al., 2014)	0.73-0.76[c]	~6.76×10^{-13}
Mouse	~5600	~0.0036(Cayrou et al., 2011)	0.79[d] (Ahuja et al., 2016)	~5.03×10^{-13}

a – chromosome missegregation rate; b – deleterious mutation rate; c – rate of cells with UFBs and 53BP1 lesions in mitosis and G1 phase respectively in U2-OS cell line; d - rate of cells with 53BP1 lesions in G1 in MEF cell line.

However in transcriptionally silent early embryos $P(\text{error})$ is minimized by genomic saturation with ROs and thus ρ is close to 1. Hence, for early embryos Eq. ($R9$) takes the very simple form $P(\text{error}) = UN_g$ and the total observed error rate of embryo death $P_{observed} \geq P(\text{error})$ and thus

$$\frac{P_{observed}}{N_g} \geq U \qquad (R11).$$

In early embryos, the only functional genomic activity is replication; hence replication error at these stages would substantially contribute to the embryonic mortality rate, and thus the $P_{observed}$ would not be far from the $P(\text{error})$ but still there could be other internal and external causes that might induce embryonic mortality. We estimate the error rate per cell division in early embryos from the normal survival rate during early embryogenesis and using this as $P_{observed}$ we can verify the inequality in Eq. ($R11$). $P_{observed}$ values from different organisms from very different phyla are in excellent agreement to Eq. ($R11$) (Table 2).

Table 2: $P_{observed}$ values in different embryos confirm the constant U.

Organism	N_g (Mbp)	Embryo stage[a]	Cell number	Survival (%)	$P_{observed}$	$\frac{P_{observed}}{N_g}$
Human	~6000	blastocyst	64-107(Eakin and Behringer, 2004)	73-78(Westphal et al., 2003)	~0.00333	~5.55×10^{-13}
				49.8(Thomas et al., 2010)	~0.00807	~13.5×10^{-13}
				42(Hardy et al., 1989)	~0.01004	~16.7×10^{-13}
Mice	~5600	blastocyst	73.7 ± 7.1(Lin et al., 2003)	81.9(Lin et al., 2003)	~0.00269	~4.81×10^{-13}
Rabbit	~5600	blastocyst	~128(Sultana et al., 2009)	73-78(Sultana et al., 2009)	~0.00224	~4.01×10^{-13}
Bovine	~6000	day 7 blastocyst	162 ± 60(Leidenfrost et al., 2011)	65.2(LOPATÁROVÁ et al., 2001)	~0.00269	~4.48×10^{-13}
				25-27(Alomar et al., 2008; Dovolou et al., 2014; LOPATÁROVÁ et al., 2001; Wang et al., 2014)	~0.00815	~13.6×10^{-13}
Horse	~5400	morula	16-32("Equus caballus, horse: embryology, life cycle and developmental stages at Geochembio, " n.d.)	91(Ball et al., 1989)	~0.00392	~7.26×10^{-13}
C. elegans	~200	neonate	~1000	88-89(Hsu et al., 2012)	~ 0.000117	~5.83×10^{-13}
Bread Wheat	~ 102000	globular	~10-20(Bakos et al., 2009)	0.51-0.66(Bakos et al., 2009)	~0.035	~3.43×10^{-13}

a – according to the stage from which the data was obtained

5.2.2 Maximum genome length in eukaryotic life

By setting $P(\text{error}) = 1$ in most extreme condition of a genome completely saturated with origins where $\rho = 1$, in Eq. ($R9$) together with $U = 3.33 \times 10^{-13}$, we get the maximum possible DNA content in a cell that can be replicated in a single cell cycle as 3.00×10^{12} bp. In *Drosophila* syncytium, rapid nuclear division stops at 13th cycle (Telley et al., 2012). Diploid drosophila genome is ~350 Mbp long and ~8000 nuclei at 13th cycle of nuclear division constitutes a total DNA content 2.8×10^{12} bp in the syncytial cell. Hence due to the upper limit constraint from U on DNA amount that can be replicated does not permit the next cycle of nuclear division and after 13th cycle individual nuclei start to cellularize at mid-blastula transition (MBT). Haploid mutant of drosophila with half the genomic DNA (~175 Mbp) showed 14th nuclear division before cellularization (Edgar et al., 1986) where ~16000 nuclei in the 14th cycle makes total DNA content in the syncytium similar to that of the diploid drosophila embryos 13th nuclear division cycle.

Moreover, in chicken and zebrafish embryos, during early embryogenesis rapid cleavage produces a balstoderm containing a mound of cells that have open connection to each other inside the large yolk cell. MBT starts after rapid cleavage is stopped at 10th cycle with ~1000 cell-nuclei making the total DNA content of the yolk cell embryo as 2.4×10^{12} bp and 2.8×10^{12} bp respectively (Gilbert, 2000; Kane and Kimmel, 1993; Nagai et al., 2015; Sheng, 2014). Together these data suggest, the maximum replicable DNA content in a single cycle of replication is determined by the constant U, and indeed it imposes a strong constraint over the course of embryogenesis in organisms that have syncytial or syncytium-like (openly connected cell mass) developmental stages. The maximum known genome length, found in *Amoeba dubia*, is ~0.67×10^{12} bp long, while the octaploid form of maximum known

genome, in *Paris japonica*, makes the total genomic content as ~1.2×10^{12} bp. These data strongly corroborate the functional biological significance of the constant U that imposes the constraint of maximum replicable DNA content of 3×10^{12} bp.

5.3 Discussion

The clear agreement of $U_{calculated}$ and $\frac{P_{observed}}{N_g}$ values to the constant U strongly demonstrates that the proposed constant is actively functional in biology. Intuitively this is understood from the definition of U as it is the ratio between two molecular biologically conserved quantities i.e. nucleosome length and fork stalling rate per nucleotide, hence any cell that replicates DNA packaged in nucleosomes with machineries found conserved across eukaryotes should comply to U. The constant U is independent of genomic or organismic complexity. Fluctuations in nucleosome length due to chromatin remodeling and intrinsic variability in linker space does not affect U as it only considers the minimum possible separation between two adjacent ROs which cannot be less than ~200 bp due to the DNA wrapped histone core and footprints of origin licensing molecules. Genome wide variation in fork stalling distance is tiny as it is much larger than the individual inter-RO gaps, also it has been shown that decreased fork stalling distance increases the error rate while increase in fork stalling distance in larger genomes do not help to minimize the error (Al Mamum et al., 2016).

Deviation from the constant U could have different implications to the replication fidelity. Much larger $U_{calculated}$ would signify the increased noise in the system, which can be interpreted as increased replication error. In cancer cells, increased chromosomal instability is a common feature, which is also related to increased

replication error. Hence, our prediction in the context of U would be frequency of markers for unreplicated DNA such as 53BP1 and UFBs in transformed or tumorigenic cells is higher than that is observed in non-transformed or normal somatic cells. This could be tested directly by experiments to quantify those molecular markers during cell division. On the other hand, if the $U_{calculated}$ is smaller than the standard value of $U = 3.33 \times 10^{-13}$, this would indicate either for replication machinery different from eukaryotes which has much smaller fork-stalling rate or the DNA organization is different i.e. the nucleosome core could be smaller in the known biological context. In case of embryonic development, U defines a default mortality for a given genome length by means of replication error and this should be the minimal mortality rate of early transcriptionally quiescent embryos in absolute absence of any other internal, external or environmental death factors. Due to the additional cues from such beyond-only-replication-error the $P_{observed}$ could be higher than the default $P(\text{error})$ but deviations in this default fixed by the organisms' genome length would imply direct threat for growth and development. The simplest prediction from our model in this context would be the observed variation in the embryonic survival data (e.g. human and bovine) is reflective of the variability in replication errors which might be the consequence of different factors like replicative stress caused by the culture media or the intrinsic factors like ageing of the oocyte.

Replication fidelity in different organisms with different genome lengths would be also influenced by this 'universal replication constant'. As U sets 'default minima' for replication error depending on the organisms' genome size, these minima would significantly drive the strategies employed by different organisms to maintain replication fidelity. Hence, organisms with small genome has small 'default minima' and thus they also have simpler strategies to confirm faithful replication, like different

yeasts only maintain a statistically regular RO spacing by means of defined RO loading sites on their genome and successfully maintains a negligible replication error rate. By contrast, constant U sets greater 'default minima' for larger genomes and organisms require additional complex mechanisms to ensure higher rate of replication error is compensated with efficient error corrections as we discussed in chapter 4 for higher eukaryotes. Increasingly larger genome length would imply greater replicative complexity, which might have a strong impact on other biological functions like developmental and organizational complexity. Thus, there has to be a trade off between organismic complexity and replicative complexity and this trade off might explain why single cell amoeba has ~100 fold larger genomic content than human.

Chapter 6

Conclusion and future directions

6.1 Discussion

In this thesis, we have presented an example of the applied simple model strategy that we discussed in the introductory chapter. Based on the study of DNA replication biology and occasional irreversible replication fork (RF) failures, the very simple assumption that each of the DNA base replicated by the polymerase has an intrinsic, independent and tiny probability for irreversible RF stalling was made. This is followed by the conceptual model of equally spaced ROs over the genome would minimize the error from double-fork stalls (DFSs) in the bulk of the chromosomes while putting replication origins (ROs) very close to the telomeric ends would minimize the damage from telomeric-fork stalls (TFSs).

Mathematical formalization of these concepts, using probability theory and statistics, provided the formulas for both DFSs and TFSs, Eq. (*R*1) and (*R*2) in chapter 3. For application, these formulas require the RO positions that are available from the whole genome RO mapping studies in order to calculate the probability of replication errors in a given genome. Hence, the formal model connects the experimental biology to the primary simple concept and together provides a result that is used to verify the model. We showed in chapter 3 that the faithful complete replication of yeast genome is maximized with regular RO distribution imposing a statistical limit to the maximum inter-RO gap in the genome as well as by putting ROs very close to the ends of the linear chromosomes. All of these theoretical predictions are validated with RO mapping data from five different yeast species. By equalizing the DFS and TFS

probability, concept originates from the basic idea that the end and bulk error need to be balanced for risk minimization, we predicted the value for median fork stalling distance N_s as ~12 Mbp which is again in very good agreement to the experimentally measured value of this parameter. Ultimately the model revealed that in yeasts indeed the ROs are significantly biased for the simplest tactic of equally spacing the ROs to minimize the RF stalling errors, and additionally we had the prediction for higher eukaryotes with longer genomes that they must possess extra-safeguards to confirm replication fidelity.

We extended our mathematical model in chapter 4 to capture the increased variability in RO spacing in higher eukaryotes due to their much larger genomes. Genome length changes from Megabases in yeasts to Gigabases in higher eukaryotes. Along with the genome length transition, we showed that the regularity in RO distribution is lost and DFS errors become increasingly inevitable in genomes spanning ~100 Megabases to ~10 Gigabases, but occurrence of errors is low i.e. less than three, as well as the larger inter-RO gaps dominate the error probability in larger genomes while the largest gap is constrained by N_s. All of these predictions are tested against RO mapping data from yeast, *Arabidopsis*, *Drosophila* and human. We showed theoretical distribution of error events agrees nicely to the experimentally measured post-replicative error markers involved with error correction in human cell lines i.e. 53BP1 nuclear bodies and UFBs and these data again confirmed N_s as ~12 Mbp.

Median fork stalling distance N_s is the reciprocal of the per nucleotide fork stall rate q and $N_s \approx 12$ Mbp provides $q \approx 5.8\times10^{-08}$, due to the conserved nature of replication machinery across eukaryotes this value is conserved too. And, in a genome organized in nucleosomes with histone cores wrapped with 147 bp of DNA, the

minimum possible distance between two adjacent ROs, N_{nuc}, is the sum of the wrapped DNA and the length of footprints of RO-licensing factors i.e. MCM2-7 double hexamer which has an effective footprint of ~60 bp. Taking these two conserved molecular factors together i.e. $q \approx 5.8\times10^{-08}$ and $N_{nuc} \approx 200$ bp in chapter 5, we derived a universal eukaryotic replication constant, $U = \frac{1}{2}q^2N_{nuc} \approx 3.33\times 10^{-13}$ that connects the genome length to the early embryonic growth as well as to the somatic and single celled eukaryotes' replication efficiency through the molecular determinants of q and N_{nuc}. Different organisms from distant phyla confirmed the proposed universal replication constant in eukaryotes.

For all the output results from the model illustrated throughout the different sections in chapter 3, 4 and 5, the inputs to the model were only experimentally known RO positions, N_s and N_{nuc}. In the introductory chapter, we discussed the output to input ratio as a scale of simplicity and strength of a model as well as more and more parameter tuning would compromise the predictive power of a model. And throughout this thesis we presented a model which had very few inputs that are actually definite values in nature rather than adjustable parameters and produced many more predictions all of which are validated by data. All together, this has been a crucial demonstration of the significance of new idea and concepts along with the illustration of the efficiency of the simple modeling strategy. We hope and expect that this work will inspire biologists to think more conceptually with simplicity being a guide in understanding complex problems as well as theoreticians to come forward with simpler concepts to capture complex biology more extensively.

6.2 Future prospects

6.2.1 RO density could influence embryonic rapid cleavage and blastomere potency

Cleavage stage embryo does not grow in volume rather the daughter cells shares the same cytoplasmic content deposited in the zygote and the cell number with the shared cytoplasm differs from organism to organism e.g. 2-16 cells in mammals and 300-4000 cells amphibians (Gilbert, 2000). At this stage, the zygotic genome is transcriptionally inactive and the maternal mRNA deposits provide necessary functionalities to the zygote until maternal to zygotic transition (MZT), which occurs toward the end of early cleavage division cycles (Langley et al., 2014). During this transcriptional quiescence, the post-replicative error repair mechanisms are absent and the embryos resort to the 'genome saturation with ROs' strategy to minimize the replication errors arising from RF failures (Ge et al., 2015; Goldar et al., 2008; Hutchins et al., 2016). Considering the zygote to be totally saturated with ROs along with the cytoplasmic net-deposit of factors contributing to replication and cell division together synthesizes a simple model of replication efficiency of these early embryonic cells. In yolk-rich eggs as in amphibians with high maternal deposit would facilitate the embryo to go for more rounds of cleavage divisions than the yolk-less eggs as in mammals depending on the density of ROs at each cleavage cycle. In yolk-less mammalian egg, due to the geometrically decreasing RO density in daughter cells, cleavage stops very early and MZT happens soon after two to three cycles of cleavage divisions. While in amphibian eggs, the high maternal deposit of such factors keeps the RO density up and allows the embryo to go up to 12 cleavage cycles before zygotic genome becomes fully active. Similarly, like the cleavage cycles, dynamics of RO density also influence the stemness of the embryonic cells. In mouse and human the embryonic cells loses pluripotency by late morula to blastocyst stage

(~32-128 cell stage) (Condic, 2014; Suwińska, 2012; Tarkowski et al., 2010). In contrary, the *Xenopus* mid-blastula cells (~4000 cell stage) are still pluripotent (Heasman et al., 1984) and that seems to reciprocate the difference in cleavage cycles between mammals and amphibians. These observations together hint at the important role of RO density dynamics during early development of multicellular organisms. Moreover, due to replication being the only genomic activity at the early stages of embryogenesis, cell death events at these stages are perhaps the direct consequence of replication failure. Hence, such available data as in bovine (Leidenfrost et al., 2011) and other organisms can be used to investigate this proposed framework of RO density dynamic contributing significantly during early embryogenesis using our model.

6.2.2 Replication error could be a cue for stem cell differentiation

Stem cell replacement and self-renewal in both development and adult homeostasis is a classic discussion of modern biology. Several models have been proposed to explain the ability of stem cells to maintain tissue homeostasis: rare long lived quiescent cells that give rise to stem and differentiated cells through asymmetric cell division; equipotent actively dividing stem cells give rise to stem and differentiated progeny via stochastic cell fate decisions (Klein and Simons, 2011). Recent experimental observations have suggested stem cell replacement rate comparable to cell division rate in different tissues (Klein et al., 2010; Lopez-Garcia et al., 2010; Snippert et al., 2010). Also both symmetric and asymmetric divisions can occur simultaneously in the tissue systems (Clayton et al., 2007; Klein and Simons, 2011). These evidences indicate that stem cell renewal and replacement could be a regular tissue stem cell function rather than being very specialized rare events. With this prospective background, we can test our model in tissue stem cell dynamic by quantitative

prediction of replication error and distribution, detection and correction of such errors in post-replicative stages. There is now considerable evidence that unreplicated DNA as a consequence of RF failures as discussed in chapter 4, can pass through the mitosis and get resolved in the following cell cycle and molecular markers like 53BP1 nuclear bodies and UFBs can be used to experimentally quantify such error events (Ahuja et al., 2016; Al Mamum et al., 2016; Lukas et al., 2011; Moreno et al., 2016). Our model suggests in a large genome like human size (~6000 Mbp), replication errors are very likely and yet their number is low (between 0-3), hence during chromosome segregation in mitosis the unreplicated DNA in error-torn pieces has the chance of differential distribution in the daughter cells depending on the actual number of errors. If there were no error, both the daughters would be fine making a symmetric division scenario where two stem cells are produced from the mother stem cell. If there was one or few error then there is a probability that the unreplicated bits of DNA might end up in one daughter while the other is free from error making a asymmetric division scenario, and two or more errors shared in both daughter makes a symmetric division scenario where from a stem cell two differentiated cell is produced. There is evidence that stem cells do not show the post-replicative error markers in normal conditions while under replicative stress they start to show such markers which are normally observed in cultures of differentiated cells and replicative stress also increases the rate of such errors in cell culture conditions (Ahuja et al., 2016; Moreno et al., 2016). As well as the embryonic stem cells shows diminishing potency with advent of embryonic cell death that could be the result of replicative stress build-up due to decreased density of licensed ROs i.e. increased error proneness. Quantitative prediction of these error and stochastic distribution of the unreplicated DNA arising at the error in the daughter cells, hence can be tested

against the stem cell differentiation rates in lineage tracing data in different tissues and this is a nice scope for the future implication of our current model.

6.2.3 Implication of the model in *Archaea* and bacteria

The 'universal replication constant' discussed in chapter 5 is based on the two conserved molecular factors which are the per-nucleotide fork stall rate and the nucleosome length that defines the minimum possible inter-RO distance in a genome. We have discussed only the eukaryotes in this thesis and shown different organisms from different eukaryotic phyla follows the pressure imposed by the 'universal replication constant' in replication and early development as well as the different strategies to maintain replication fidelity is under influence of the constant in different scales of eukaryotic life spanning from single celled yeasts to the complex higher eukaryotes. DNA replication in *Archaea* is similar to that of eukaryotes and the *Archaeal* replication machineries are more akin to the eukaryotes than prokaryotes and specifically the structure of MCM helicases are conserved (Barry and Bell, 2006; Miller et al., 2015). As well as the histone tetramer core of the *Archaeal* nucleosomes, homolog of the eukaryotic nucleosome core histones, is wrapped with ~85 bp DNA which has comparative structural similarity to the eukaryotic nucleosome organization (Bailey and Reeve, 1999; Pereira and Reeve, 1998; Reeve et al., 2004). Qualitative replication biology in *Archaea* thus being similar to eukaryotes, hints to a similar replication constant comparable to the 'Universal replication constant' in eukaryotes. Considering the conservation of MCM structure, N_{nuc} value in *Archaea* comes to be ~140 bp (~200 bp in eukaryotes) and together with similar q value as in eukaryotes, the replication constant $U = \frac{1}{2} q^2 N_{nuc} \approx 2.33\times10^{-13}$. It is interesting that the value of U less than 3.33×10^{-13} (as calculated in eukaryotes) suggests either much smaller per-nucleotide stalling rate (which is less likely due to the homology of the replication

machinery between *Archaea* and *Eukarya*) or smaller length of DNA at nucleosome core and in effect *Archaeal* nucleosome cores are wrapped with lesser DNA bases than the eukaryotes. Data on different replicative errors like chromosomal instability rate would help to validate this prospect of the replication constant in *Archaea*. 2-10 Mbp *Archaeal* genome carries only one to three ROs. In chapter 4, we discussed that the median stalling distance N_s constrains the permissible inter-RO distance and the maximum inter-RO distance is limited by the range ~$0.5N_s$. It is worth noticing that this boundary effect is also agreeable to the *Archaeal* genomes and it might be extended to some degree to the bacterial genomes too, which are 2-8 Mbp single circular chromosomes with one single RO. The characteristic difference in DNA organization between bacteria and eukaryotes and associated biology of replication fidelity and error could open a window for application of our model in broader scope of biology spanning prokaryotes to eukaryotes.

6.2.4 Implication of the model in therapeutics

Replicative stress, characterized by RF failures, has been linked to oncogenic activation during cancer (Gorgoulis and Halazonetis, 2010; Negrini et al., 2010; Zeman and Cimprich, 2014). Observed early DNA damage responses in cancer cells with no genotoxic-therapies could be the result of this replicative stress and such damage-induced senescence is a barrier for tumorigenesis (Bartkova et al., 2006, 2005). Thus induced replicative stress by depleting the capacity of origin licensing in tumor cells could become a potential way to stop carcinogenesis at early stages of cancer development and therapeutic drugs based on this strategy are already in clinical and preclinical trials. Detail of this prospective future therapy for cancer treatment is reviewed elsewhere (Dobbelstein and Sørensen, 2015). The most important implication of our model in this context would the quantitative assessment

of replication error profile for the target cells with targeted depletion in the density of licensed ROs or vice versa. Our model has already been successfully applied to quantify replication error probabilities in the context of RO depletion and model estimates are experimentally validated in HeLa, IMR90 and U2-OS cell lines (Al Mamum et al., 2016; Moreno et al., 2016). Hence, our model has strong potential to be applied for quantitative assessment of the biological efficiency of replication stress based therapeutics in cancer.

The universal eukaryotic replication constant U directly connects the genome size to the developmental robustness during embryogenesis through the fixed 'default minima' in replication error in early transcriptionally silent cleavage stage embryos. So using our model, we can quantify the idealized embryonic mortality of a given organism in the early developmental stages. This has direct implication in in-vitro fertilization (IVF) and in-vitro maturation (IVM) technologies. Knowing the 'default minima' that only require the genome size of the organism, like for human embryo we can have quantitative estimate of the contribution of non-replicative mortality factors from the embryonic survival rates. This could be a good quantitative measure of the impact of culture media, growth technique, and could also help to measure the internal or epigenetic mortality cue like maternal ageing under the same media conditions.

6.3 Concluding remarks

In this thesis, we have presented an example of simple applied conceptual ideas and assumptions in very complex biological issues namely DNA replication across eukaryotes. Using probability theory, the model has bridged the high throughput experimental data and conceptual simplicity bringing forth clear predictions, which

are biologically validated by data. This is an illustration of a simpler conceptual approach to complex biology that has now become a necessity after decades of molecular biology, which caused a silent revolution in biological data. With big data now available on molecular details of different complex biological issues from cell biology to organismic homoestasis, new ideas and concepts need to be brought forward in order to guide our tour in this messy and massive data park. In that regard, we hope and expect the outcomes from this study would help the biological community to look at the DNA replication biology and related data bank in a broader systemic and unified context of biological simplicity and it would help shed light on the ever increasing complexity in the biological systems as we continue to discover more of the biology.

Bibliography

Abbas, T., Keaton, M.A., Dutta, A., 2013. Genomic Instability in Cancer. Cold Spring Harb. Perspect. Biol. 5, a012914. doi:10.1101/cshperspect.a012914

Agier, N., Romano, O.M., Touzain, F., Lagomarsino, M.C., Fischer, G., 2013. The Spatiotemporal Program of Replication in the Genome of Lachancea kluyveri. Genome Biol. Evol. 5, 370–388. doi:10.1093/gbe/evt014

Aguilera, A., Gómez-González, B., 2008. Genome instability: a mechanistic view of its causes and consequences. Nat. Rev. Genet. 9, 204–217. doi:10.1038/nrg2268

Ahuja, A.K., Jodkowska, K., Teloni, F., Bizard, A.H., Zellweger, R., Herrador, R., Ortega, S., Hickson, I.D., Altmeyer, M., Mendez, J., Lopes, M., 2016. A short G1 phase imposes constitutive replication stress and fork remodelling in mouse embryonic stem cells. Nat. Commun. 7, 10660. doi:10.1038/ncomms10660

Albergante, L., Blow, J.J., Newman, T.J., 2014. Buffered Qualitative Stability explains the robustness and evolvability of transcriptional networks. eLife 3, e02863. doi:10.7554/eLife.02863

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2002. The Universal Features of Cells on Earth, in: Molecular Biology of the Cell. Garland Science, New York.

Allen, C., Ashley, A.K., Hromas, R., Nickoloff, J.A., 2011. More forks on the road to replication stress recovery. J. Mol. Cell Biol. 3, 4–12. doi:10.1093/jmcb/mjq049

Al Mamum, M., Albergante, L., Moreno, A., Blow, J.J., Newman, T.J., 2016. Inevitability and containment of replication errors for eukaryotic genome lengths spanning Megabase to Gigabase. PNAS.

Alomar, M., Tasiaux, H., Remacle, S., George, F., Paul, D., Donnay, I., 2008. Kinetics of fertilization and development, and sex ratio of bovine embryos produced using the semen of different bulls. Anim. Reprod. Sci. 107, 48–61. doi:10.1016/j.anireprosci.2007.06.009

Alver, R.C., Chadha, G.S., Blow, J.J., 2014a. The contribution of dormant origins to genome stability: From cell biology to human genetics. DNA Repair, Cutting-edge Perspectives in Genomic Maintenance 19, 182–189. doi:10.1016/j.dnarep.2014.03.012

Alver, R.C., Zhang, T., Josephrajan, A., Fultz, B.L., Hendrix, C.J., Das-Bradoo, S., Bielinsky, A.-K., 2014b. The N-terminus of Mcm10 is important for interaction with the 9-1-1 clamp and in resistance to DNA damage. Nucleic Acids Res. 42, 8389–8404. doi:10.1093/nar/gku479

Aparicio, O.M., 2013. Location, location, location: it's all in the timing for replication origins. Genes Dev. 27, 117–128. doi:10.1101/gad.209999.112

Arias, E.E., Walter, J.C., 2007. Strength in numbers: preventing rereplication via multiple mechanisms in eukaryotic cells. Genes Dev. 21, 497–518. doi:10.1101/gad.1508907

Bailey, K.A., Reeve, J.N., 1999. DNA repeats and archaeal nucleosome positioning. Res. Microbiol. 150, 701–709. doi:10.1016/S0923-2508(99)00122-9

Bakos, F., Szabó, L., Olmedilla, A., Barnabás, B., 2009. Histological comparison between wheat embryos developing in vitro from isolated zygotes and those

developing in vivo. Sex. Plant Reprod. 22, 15–25. doi:10.1007/s00497-008-0087-7

Ball, B.A., Little, T.V., Weber, J.A., Woods, G.L., 1989. Survival of day-4 embryos from young, normal mares and aged, subfertile mares after transfer to normal recipient mares. J. Reprod. Fertil. 85, 187–194.

Barberis, M., Spiesser, T.W., Klipp, E., 2010. Replication Origins and Timing of Temporal Replication in Budding Yeast: How to Solve the Conundrum? Curr. Genomics 11, 199–211. doi:10.2174/138920210791110942

Barry, E.R., Bell, S.D., 2006. DNA Replication in the Archaea. Microbiol. Mol. Biol. Rev. 70, 876–887. doi:10.1128/MMBR.00029-06

Bartkova, J., Horejsí, Z., Koed, K., Krämer, A., Tort, F., Zieger, K., Guldberg, P., Sehested, M., Nesland, J.M., Lukas, C., Ørntoft, T., Lukas, J., Bartek, J., 2005. DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. Nature 434, 864–870. doi:10.1038/nature03482

Bartkova, J., Rezaei, N., Liontos, M., Karakaidos, P., Kletsas, D., Issaeva, N., Vassiliou, L.-V.F., Kolettas, E., Niforou, K., Zoumpourlis, V.C., Takaoka, M., Nakagawa, H., Tort, F., Fugger, K., Johansson, F., Sehested, M., Andersen, C.L., Dyrskjot, L., Ørntoft, T., Lukas, J., Kittas, C., Helleday, T., Halazonetis, T.D., Bartek, J., Gorgoulis, V.G., 2006. Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. Nature 444, 633–637. doi:10.1038/nature05268

Bebenek, A., 2008. [DNA replication fidelity]. Postepy Biochem. 54, 43–56.

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., Lemaitre, J.-M., 2012. Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-

quadruplex consensus motifs. Nat. Struct. Mol. Biol. 19, 837–844. doi:10.1038/nsmb.2339

Blow, J.J., Dutta, A., 2005. Preventing re-replication of chromosomal DNA. Nat. Rev. Mol. Cell Biol. 6, 476–486. doi:10.1038/nrm1663

Blow, J.J., Ge, X.Q., 2009. A model for DNA replication showing how dormant origins safeguard against replication fork failure. EMBO Rep. 10, 406–412. doi:10.1038/embor.2009.5

Blow, J.J., Ge, X.Q., Jackson, D.A., 2011. How dormant origins promote complete genome replication. Trends Biochem. Sci. 36, 405–414. doi:10.1016/j.tibs.2011.05.002

Blow, J.J., Hodgson, B., 2002. Replication licensing — Origin licensing: defining the proliferative state? Trends Cell Biol. 12, 72–78. doi:10.1016/S0962-8924(01)02203-6

Bohgaki, T., Bohgaki, M., Hakem, R., 2010. DNA double-strand break signaling and human disorders. Genome Integr. 1, 15. doi:10.1186/2041-9414-1-15

Borowiec, J.A., Schildkraut, C.L., 2011. Open sesame: activating dormant replication origins in the mouse immunoglobulin heavy chain (Igh) locus. Curr. Opin. Cell Biol. 23, 284–292. doi:10.1016/j.ceb.2011.04.004

Brümmer, A., Salazar, C., Zinzalla, V., Alberghina, L., Höfer, T., 2010. Mathematical Modelling of DNA Replication Reveals a Trade-off between Coherence of Origin Activation and Robustness against Rereplication. PLoS Comput Biol 6, e1000783. doi:10.1371/journal.pcbi.1000783

Burgers, P.M.J., 2009. Polymerase dynamics at the eukaryotic DNA replication fork. J. Biol. Chem. 284, 4041–4045. doi:10.1074/jbc.R800062200

Cayrou, C., Coulombe, P., Vigneron, A., Stanojcic, S., Ganier, O., Peiffer, I., Rivals, E., Puy, A., Laurent-Chabalier, S., Desprat, R., Méchali, M., 2011. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res. 21, 1438–1449. doi:10.1101/gr.121830.111

Cherry, J.M., 2015. The Saccharomyces Genome Database: A Tool for Discovery. Cold Spring Harb. Protoc. 2015, pdb.top083840. doi:10.1101/pdb.top083840

Cisneros, L.H., Newman, T.J., 2014. Quantifying metastatic inefficiency: rare genotypes versus rare dynamics. Phys. Biol. 11, 046003. doi:10.1088/1478-3975/11/4/046003

Clayton, E., Doupé, D.P., Klein, A.M., Winton, D.J., Simons, B.D., Jones, P.H., 2007. A single type of progenitor cell maintains normal epidermis. Nature 446, 185–189. doi:10.1038/nature05574

Cobb, J.A., Schleker, T., Rojas, V., Bjergbaek, L., Tercero, J.A., Gasser, S.M., 2005. Replisome instability, fork collapse, and gross chromosomal rearrangements arise synergistically from Mec1 kinase and RecQ helicase mutations. Genes Dev. 19, 3055–3069. doi:10.1101/gad.361805

Condic, M.L., 2014. Totipotency: What It Is and What It Is Not. Stem Cells Dev. 23, 796–812. doi:10.1089/scd.2013.0364

Cooper, G.M., 2000. The Eukaryotic Cell Cycle, in: The Cell: A Molecular Approach. 2nd Edition. Sunderland (MA): Sinauer Associates, Boston University.

Costas, C., de la Paz Sanchez, M., Stroud, H., Yu, Y., Oliveros, J.C., Feng, S., Benguria, A., López-Vidriero, I., Zhang, X., Solano, R., Jacobsen, S.E., Gutierrez, C., 2011. Genome-wide mapping of Arabidopsis thaliana origins of

DNA replication and their associated epigenetic marks. Nat. Struct. Mol. Biol. 18, 395–400. doi:10.1038/nsmb.1988

Cotobal, C., Segurado, M., Antequera, F., 2010. Structural diversity and dynamics of genomic replication origins in Schizosaccharomyces pombe. EMBO J. 29, 934–942. doi:10.1038/emboj.2009.411

Cowan, R., 2003. Stochastic models for DNA replication. North Holland.

Curtin, N., Sharma, R., 2015. PARP Inhibitors for Cancer Therapy. Humana Press.

Debatisse, M., Le Tallec, B., Letessier, A., Dutrillaux, B., Brison, O., 2012. Common fragile sites: mechanisms of instability revisited. Trends Genet. TIG 28, 22–32. doi:10.1016/j.tig.2011.10.003

Deniz, Ö., Flores, O., Aldea, M., Soler-López, M., Orozco, M., 2016. Nucleosome architecture throughout the cell cycle. Sci. Rep. 6. doi:10.1038/srep19729

De Piccoli, G., Katou, Y., Itoh, T., Nakato, R., Shirahige, K., Labib, K., 2012. Replisome stability at defective DNA replication forks is independent of S phase checkpoint kinases. Mol. Cell 45, 696–704. doi:10.1016/j.molcel.2012.01.007

Dirac, P.A.M., 1982. The principles of quantum mechanics. Oxford University Press, USA.

Di Rienzi, S.C., Lindstrom, K.C., Lancaster, R., Rolczynski, L., Raghuraman, M.K., Brewer, B.J., 2011. Genetic, genomic, and molecular tools for studying the protoploid yeast, L. waltii. Yeast Chichester Engl. 28, 137–151. doi:10.1002/yea.1826

Dobbelstein, M., Sørensen, C.S., 2015. Exploiting replicative stress to treat cancer. Nat. Rev. Drug Discov. 14, 405–423. doi:10.1038/nrd4553

Dovolou, E., Periquesta, E., Messinis, I.E., Tsiligianni, T., Dafopoulos, K., Gutierrez-Adan, A., Amiridis, G.S., 2014. Daily supplementation with ghrelin improves in vitro bovine blastocysts formation rate and alters gene expression related to embryo quality. Theriogenology 81, 565–571. doi:10.1016/j.theriogenology.2013.11.009

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.-M., Beyne, E., Bleykasten, C., Boisramé, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.-M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.-F., Straub, M.-L., Suleau, A., Swennen, D., Tekaia, F., Wésolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., Souciet, J.-L., 2004. Genome evolution in yeasts. Nature 430, 35–44. doi:10.1038/nature02579

Eakin, G.S., Behringer, R.R., 2004. Gastrulation in other mammals and humans, in: Gastrulation: From Cells to Embryo. pp. 275–287.

Eaton, M.L., Galani, K., Kang, S., Bell, S.P., MacAlpine, D.M., 2010. Conserved nucleosome positioning defines replication origins. Genes Dev. 24, 748–753. doi:10.1101/gad.1913210

Edgar, B.A., Kiehle, C.P., Schubiger, G., 1986. Cell cycle control by the nucleo-cytoplasmic ratio in early Drosophila development. Cell 44, 365–372.

Equus caballus, horse: embryology, life cycle and developmental stages at Geochembio [WWW Document], n.d. URL http://www.geochembio.com/biology/organisms/horse/horse-life-cycle-and-development.html (accessed 3.16.16).

Everitt, B.S., 1998. The Cambridge Dictionary of Statistics. Cambridge University Press, Cambridge, UK.

Evrin, C., Clarke, P., Zech, J., Lurz, R., Sun, J., Uhle, S., Li, H., Stillman, B., Speck, C., 2009. A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. Proc. Natl. Acad. Sci. 106, 20240–20245. doi:10.1073/pnas.0911500106

Forster, M.R., 2001. "The New Science of Simplicity," in Arnold Zellner, Hugo Keuzenkamp, and Michael McAleer (eds.), Simplicity, Inference and Modelling. pp. 83–117.

Francis, D., Davies, M.S., Barlow, P.W., 2008. A strong nucleotypic effect on the cell cycle regardless of ploidy level. Ann. Bot. 101, 747–757. doi:10.1093/aob/mcn038

Gauthier, M.G., Norio, P., Bechhoefer, J., 2012. Modeling inhomogeneous DNA replication kinetics. PloS One 7, e32053. doi:10.1371/journal.pone.0032053

Gauthier, M., Herrick, J., Bechhoefer, J., 2010. Defects and DNA replication. Presented at the APS Meeting Abstracts, p. 13013.

Geiringer, H., 1940. A Generalization of the Law of Large Numbers. Ann. Math. Stat. 11, 393–401. doi:10.1214/aoms/1177731826

Ge, X.Q., Han, J., Cheng, E.-C., Yamaguchi, S., Shima, N., Thomas, J.-L., Lin, H., 2015. Embryonic Stem Cells License a High Level of Dormant Origins to Protect the Genome against Replication Stress. Stem Cell Rep. 5, 185–194. doi:10.1016/j.stemcr.2015.06.002

Ge, X.Q., Jackson, D.A., Blow, J.J., 2007. Dormant origins licensed by excess Mcm2-7 are required for human cells to survive replicative stress. Genes Dev. 21, 3331–3341. doi:10.1101/gad.457807

Ghosal, G., Chen, J., 2013. DNA damage tolerance: a double-edged sword guarding the genome. Transl. Cancer Res. 2, 107–129.

Gilbert, D.M., 2012. Replication origins run (ultra) deep. Nat. Struct. Mol. Biol. 19, 740–742. doi:10.1038/nsmb.2352

Gilbert, S.F., 2000. Early Development in Fish, in: Developmental Biology. Sinauer Associates, Sunderland (MA).

Gilbert, S.F., 2000. An Introduction to Early Developmental Processes.

Goldar, A., Labit, H., Marheineke, K., Hyrien, O., 2008. A Dynamic Stochastic Model for DNA Replication Initiation in Early Embryos. PLOS ONE 3, e2919. doi:10.1371/journal.pone.0002919

Gorgoulis, V.G., Halazonetis, T.D., 2010. Oncogene-induced senescence: the bright and dark side of the response. Curr. Opin. Cell Biol. 22, 816–827. doi:10.1016/j.ceb.2010.07.013

Haight, F.A., 1967. A Handbook of the Poisson Distribution. J. Oper. Res. Soc. - J OPER RES SOC 18.

Halazonetis, T.D., Gorgoulis, V.G., Bartek, J., 2008. An oncogene-induced DNA damage model for cancer development. Science 319, 1352–1355. doi:10.1126/science.1140735

Hardy, K., Handyside, A.H., Winston, R.M., 1989. The human blastocyst: cell number, death and allocation during late preimplantation development in vitro. Development 107, 597–604.

Hayashi, M., Katou, Y., Itoh, T., Tazumi, A., Tazumi, M., Yamada, Y., Takahashi, T., Nakagawa, T., Shirahige, K., Masukata, H., 2007a. Genome-wide localization of pre-RC sites and identification of replication origins in fission yeast. EMBO J. 26, 1327–1339. doi:10.1038/sj.emboj.7601585

Hayashi, M., Katou, Y., Itoh, T., Tazumi, M., Yamada, Y., Takahashi, T., Nakagawa, T., Shirahige, K., Masukata, H., 2007b. Genome-wide localization of pre-RC sites and identification of replication origins in fission yeast. EMBO J. 26, 1327–1339. doi:10.1038/sj.emboj.7601585

Heasman, J., Wylie, C.C., Hausen, P., Smith, J.C., 1984. Fates and states of determination of single vegetal pole blastomeres of X. laevis. Cell 37, 185–194.

Hsu, P.-C.L., O'Callaghan, M., Al-Salim, N., Hurst, M.R.H., 2012. Quantum dot nanoparticles affect the reproductive system of Caenorhabditis elegans. Environ. Toxicol. Chem. SETAC 31, 2366–2374. doi:10.1002/etc.1967

Hutchins, J.R.A., Aze, A., Coulombe, P., Méchali, M., 2016. Characteristics of Metazoan DNA Replication Origins, in: Hanaoka, F., Sugasawa, K. (Eds.), DNA Replication, Recombination, and Repair. Springer Japan, pp. 23–52.

Jones, R.M., Kotsantis, P., Stewart, G.S., Groth, P., Petermann, E., 2014. BRCA2 and RAD51 Promote Double-Strand Break Formation and Cell Death in Response to Gemcitabine. Mol. Cancer Ther. 13, 2412–2421. doi:10.1158/1535-7163.MCT-13-0862

Kane, D.A., Kimmel, C.B., 1993. The zebrafish midblastula transition. Dev. Camb. Engl. 119, 447–456.

Karschau, J., Blow, J.J., de Moura, A.P.S., 2012. Optimal placement of origins for DNA replication. Phys. Rev. Lett. 108, 058101. doi:10.1103/PhysRevLett.108.058101

Klein, A.M., Nakagawa, T., Ichikawa, R., Yoshida, S., Simons, B.D., 2010. Mouse germ line stem cells undergo rapid and stochastic turnover. Cell Stem Cell 7, 214–224. doi:10.1016/j.stem.2010.05.017

Klein, A.M., Simons, B.D., 2011. Universal patterns of stem cell fate in cycling adult tissues. Development 138, 3103–3111. doi:10.1242/dev.060103

Kornberg, A., Lehman, I.R., Bessman, M.J., Simms, E.S., 1956. Enzymic synthesis of deoxyribonucleic acid. Biochim. Biophys. Acta 21, 197–198. doi:10.1016/0006-3002(56)90127-5

Langley, A.R., Smith, J.C., Stemple, D.L., Harvey, S.A., 2014. New insights into the maternal to zygotic transition. Development 141, 3834–3841. doi:10.1242/dev.102368

Leidenfrost, S., Boelhauve, M., Reichenbach, M., Güngör, T., Reichenbach, H.-D., Sinowatz, F., Wolf, E., Habermann, F.A., 2011. Cell arrest and cell death in mammalian preimplantation development: lessons from the bovine model. PloS One 6, e22121. doi:10.1371/journal.pone.0022121

Leman, A.R., Noguchi, E., 2013. The Replication Fork: Understanding the Eukaryotic Replication Machinery and the Challenges to Genome Duplication. Genes 4, 1–32. doi:10.3390/genes4010001

Leonard, A.C., Méchali, M., 2013. DNA Replication Origins. Cold Spring Harb. Perspect. Biol. 5, a010116. doi:10.1101/cshperspect.a010116

Liachko, I., Bhaskar, A., Lee, C., Chung, S.C.C., Tye, B.-K., Keich, U., 2010. A Comprehensive Genome-Wide Map of Autonomously Replicating Sequences in a Naive Genome. PLoS Genet 6, e1000946. doi:10.1371/journal.pgen.1000946

Li, H., Mitchell, J.R., Hasty, P., 2008. DNA double-strand breaks: a potential causative factor for mammalian aging? Mech. Ageing Dev. 129, 416–424. doi:10.1016/j.mad.2008.02.002

Lin, T.-C., Yen, J.-M., Gong, K.-B., Hsu, T.-T., Chen, L.-R., 2003. IGF-1/IGFBP-1 increases blastocyst formation and total blastocyst cell number in mouse embryo culture and facilitates the establishment of a stem-cell line. BMC Cell Biol. 4, 14. doi:10.1186/1471-2121-4-14

LOPATÁROVÁ, M., HOLY, L., VINKLER, A., 2001. Effect on survival of micromanipulating the zona pellucida of bovine embryos. Acta Vet. Brno 70, 49–56.

Lopez-Garcia, C., Klein, A.M., Simons, B.D., Winton, D.J., 2010. Intestinal stem cell replacement follows a pattern of neutral drift. Science 330, 822–825. doi:10.1126/science.1196236

Lukas, C., Savic, V., Bekker-Jensen, S., Doil, C., Neumann, B., Pedersen, R.S., Grøfte, M., Chan, K.L., Hickson, I.D., Bartek, J., Lukas, J., 2011. 53BP1 nuclear bodies form around DNA lesions generated by mitotic transmission of chromosomes under replication stress. Nat. Cell Biol. 13, 243–253. doi:10.1038/ncb2201

Lygeros, J., Koutroumpas, K., Dimopoulos, S., Legouras, I., Kouretas, P., Heichinger, C., Nurse, P., Lygerou, Z., 2008. Stochastic hybrid modeling of DNA

replication across a complete genome. Proc. Natl. Acad. Sci. 105, 12295–12300. doi:10.1073/pnas.0805549105

Macheret, M., Halazonetis, T.D., 2015. DNA Replication Stress as a Hallmark of Cancer. Annu. Rev. Pathol. Mech. Dis. 10, 425–448. doi:10.1146/annurev-pathol-012414-040424

Mahbubani, H.M., Chong, J.P.J., Chevalier, S., Thömmes, P., Blow, J.J., 1997. Cell Cycle Regulation of the Replication Licensing System: Involvement of a Cdk-dependent Inhibitor. J. Cell Biol. 136, 125–135.

Maya-Mendoza, A., Petermann, E., Gillespie, D.A.F., Caldecott, K.W., Jackson, D.A., 2007. Chk1 regulates the density of active replication origins during the vertebrate S phase. EMBO J. 26, 2719–2731. doi:10.1038/sj.emboj.7601714

Mazouzi, A., Velimezi, G., Loizou, J.I., 2014. DNA replication stress: Causes, resolution and disease. Exp. Cell Res., DNA DAMAGE AND REPAIR 329, 85–93. doi:10.1016/j.yexcr.2014.09.030

McIntosh, D., Blow, J.J., 2012. Dormant origins, the licensing checkpoint, and the response to replicative stresses. Cold Spring Harb. Perspect. Biol. 4. doi:10.1101/cshperspect.a012955

Méchali, M., 2010. Eukaryotic DNA replication origins: many choices for appropriate answers. Nat. Rev. Mol. Cell Biol. 11, 728–738. doi:10.1038/nrm2976

Meselson, M., Stahl, F.W., 1958. THE REPLICATION OF DNA IN ESCHERICHIA COLI*. Proc. Natl. Acad. Sci. U. S. A. 44, 671–682.

Miller, J.M., Enemark, E.J., Miller, J.M., Enemark, E.J., 2015. Archaeal MCM Proteins as an Analog for the Eukaryotic Mcm2–7 Helicase to Reveal Essential Features of Structure and Function, Archaeal MCM Proteins as an Analog for the Eukaryotic Mcm2–7 Helicase to Reveal Essential

Features of Structure and Function. Archaea Archaea 2015, 2015, e305497. doi:10.1155/2015/305497, 10.1155/2015/305497

Moreno, A., Carrington, J.T., Albergante, L., Mamun, M.A., Haagensen, E.J., Gourgolis, V.G., Newman, T.J., Blow, J.J., 2016. Unreplicated DNA remaining from unperturbed S phases passes through mitosis for resolution in daughter cells. Submitted.

Müller, C.A., Nieduszynski, C.A., 2012. Conservation of replication timing reveals global and local regulation of replication origin activity. Genome Res. 22, 1953–1962. doi:10.1101/gr.139477.112

Nagai, H., Sezaki, M., Kakiguchi, K., Nakaya, Y., Lee, H.C., Ladher, R., Sasanami, T., Han, J.Y., Yonemura, S., Sheng, G., 2015. Cellular analysis of cleavage-stage chick embryos reveals hidden conservation in vertebrate early development. Dev. Camb. Engl. 142, 1279–1286. doi:10.1242/dev.118604

Nakajima, T., 2013. Probability in biology: Overview of a comprehensive theory of probability in living systems. Prog. Biophys. Mol. Biol., Can Biology Create a Profoundly New Mathematics and Computation?Special Theme Issue on Integral Biomathics 113, 67–79. doi:10.1016/j.pbiomolbio.2013.03.007

Negrini, S., Gorgoulis, V.G., Halazonetis, T.D., 2010. Genomic instability--an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol. 11, 220–228. doi:10.1038/nrm2858

Newman, T., 2015. Biology is simple. Phys. Biol. 12, 063002. doi:10.1088/1478-3975/12/6/063002

Newman, T.J., Mamun, M.A., Nieduszynski, C.A., Blow, J.J., 2013. Replisome stall events have shaped the distribution of replication origins in the genomes of yeasts. Nucleic Acids Res. 41, 9705–9718. doi:10.1093/nar/gkt728

Nieduszynski, C.A., Blow, J.J., Donaldson, A.D., 2005. The requirement of yeast replication origins for pre-replication complex proteins is modulated by transcription. Nucleic Acids Res. 33, 2410–2420. doi:10.1093/nar/gki539

Nieduszynski, C.A., Knox, Y., Donaldson, A.D., 2006. Genome-wide identification of replication origins in yeast by comparative genomics. Genes Dev. 20, 1874–1879. doi:10.1101/gad.385306

Patel, P.K., Arcangioli, B., Baker, S.P., Bensimon, A., Rhind, N., 2006. DNA Replication Origins Fire Stochastically in Fission Yeast. Mol. Biol. Cell 17, 308–316. doi:10.1091/mbc.E05-07-0657

Pereira, S.L., Reeve, J.N., 1998. Histones and nucleosomes in Archaea and Eukarya: a comparative analysis. Extrem. Life Extreme Cond. 2, 141–148.

Picard, F., Cadoret, J.-C., Audit, B., Arneodo, A., Alberti, A., Battail, C., Duret, L., Prioleau, M.-N., 2014. The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. PLoS Genet 10, e1004282. doi:10.1371/journal.pgen.1004282

Reeve, J.N., Bailey, K.A., Li, W.–., Marc, F., Sandman, K., Soares, D.J., 2004. Archaeal histones: structures, stability and DNA binding. Biochem. Soc. Trans. 32, 227–230. doi:10.1042/bst0320227

Remus, D., Beuron, F., Tolun, G., Griffith, J.D., Morris, E.P., Diffley, J.F.X., 2009. Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. Cell 139, 719–730. doi:10.1016/j.cell.2009.10.015

Rienzi, S.C.D., Lindstrom, K.C., Mann, T., Noble, W.S., Raghuraman, M.K., Brewer, B.J., 2012. Maintaining replication origins in the face of genomic change. Genome Res. 22, 1940–1952. doi:10.1101/gr.138248.112

Riera, A., Tognetti, S., Speck, C., 2014. Helicase loading: how to build a MCM2-7 double-hexamer. Semin. Cell Dev. Biol. 30, 104–109. doi:10.1016/j.semcdb.2014.03.008

Sclafani, R.A., Holzen, T.M., 2007. Cell Cycle Regulation of DNA Replication. Annu. Rev. Genet. 41, 237–280. doi:10.1146/annurev.genet.41.110306.130308

Shalm, L.K., Meyer-Scott, E., Christensen, B.G., Bierhorst, P., Wayne, M.A., Stevens, M.J., Gerrits, T., Glancy, S., Hamel, D.R., Allman, M.S., Coakley, K.J., Dyer, S.D., Hodge, C., Lita, A.E., Verma, V.B., Lambrocco, C., Tortorici, E., Migdall, A.L., Zhang, Y., Kumor, D.R., Farr, W.H., Marsili, F., Shaw, M.D., Stern, J.A., Abellán, C., Amaya, W., Pruneri, V., Jennewein, T., Mitchell, M.W., Kwiat, P.G., Bienfang, J.C., Mirin, R.P., Knill, E., Nam, S.W., 2015. Strong Loophole-Free Test of Local Realism*. Phys. Rev. Lett. 115, 250402. doi:10.1103/PhysRevLett.115.250402

Sheng, G., 2014. Day-1 chick development. Dev. Dyn. 243, 357–367. doi:10.1002/dvdy.24087

Sherman, D.J., Martin, T., Nikolski, M., Cayla, C., Souciet, J.-L., Durrens, P., Génolevures Consortium, 2009. Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. Nucleic Acids Res. 37, D550–554. doi:10.1093/nar/gkn859

Siow, C.C., Nieduszynska, S.R., Müller, C.A., Nieduszynski, C.A., 2012. OriDB, the DNA replication origin database updated and extended. Nucleic Acids Res. 40, D682–D686. doi:10.1093/nar/gkr1091

Snippert, H.J., Flier, L.G. van der, Sato, T., Es, J.H. van, Born, M. van den, Kroon-Veenboer, C., Barker, N., Klein, A.M., Rheenen, J. van, Simons, B.D.,

Clevers, H., 2010. Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells. Cell 143, 134–144. doi:10.1016/j.cell.2010.09.016

Snyder, M., Sapolsky, R.J., Davis, R.W., 1988. Transcription interferes with elements important for chromosome maintenance in Saccharomyces cerevisiae. Mol. Cell. Biol. 8, 2184–2194.

Speck, C., Chen, Z., Li, H., Stillman, B., 2005. ATPase-dependent, cooperative binding of ORC and Cdc6p to origin DNA. Nat. Struct. Mol. Biol. 12, 965–971. doi:10.1038/nsmb1002

Spencer, F., Gerring, S.L., Connelly, C., Hieter, P., 1990. Mitotic chromosome transmission fidelity mutants in Saccharomyces cerevisiae. Genetics 124, 237–249.

Strome, E.D., Wu, X., Kimmel, M., Plon, S.E., 2008. Heterozygous screen in Saccharomyces cerevisiae identifies dosage-sensitive genes that affect chromosome stability. Genetics 178, 1193–1207. doi:10.1534/genetics.107.084103

Sultana, F., Hatori, M., Shimozawa, N., Ebisawa, T., Sankai, T., 2009. Continuous Observation of Rabbit Preimplantation Embryos In Vitro by Using a Culture Device Connected to a Microscope. J. Am. Assoc. Lab. Anim. Sci. JAALAS 48, 52–56.

Suwińska, A., 2012. Preimplantation mouse embryo: developmental fate and potency of blastomeres. Results Probl. Cell Differ. 55, 141–163. doi:10.1007/978-3-642-30406-4_8

Szerlong, H.J., Hansen, J.C., 2011. Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. Biochem. Cell Biol. Biochim. Biol. Cell. 89, 24–34. doi:10.1139/O10-139

Szilard, R.K., Jacques, P.-E., Laramée, L., Cheng, B., Galicia, S., Bataille, A.R., Yeung, M., Mendez, M., Bergeron, M., Robert, F., Durocher, D., 2010. Systematic identification of fragile sites via genome-wide location analysis of gamma-H2AX. Nat. Struct. Mol. Biol. 17, 299–305. doi:10.1038/nsmb.1754

Tarkowski, A.K., Suwińska, A., Czołowska, R., Ożdżeński, W., 2010. Individual blastomeres of 16- and 32-cell mouse embryos are able to develop into foetuses and mice. Dev. Biol. 348, 190–198. doi:10.1016/j.ydbio.2010.09.022

Telley, I.A., Gáspár, I., Ephrussi, A., Surrey, T., 2012. Aster migration determines the length scale of nuclear separation in the Drosophila syncytial embryo. J. Cell Biol. 197, 887–895. doi:10.1083/jcb.201204019

Theis, J.F., Irene, C., Dershowitz, A., Brost, R.L., Tobin, M.L., di Sanzo, F.M., Wang, J.-Y., Boone, C., Newlon, C.S., 2010. The DNA damage response pathway contributes to the stability of chromosome III derivatives lacking efficient replicators. PLoS Genet. 6, e1001227. doi:10.1371/journal.pgen.1001227

Thomas, M.R., Sparks, A.E., Ryan, G.L., Van Voorhis, B.J., 2010. Clinical predictors of human blastocyst formation and pregnancy after extended embryo culture and transfer. Fertil. Steril. 94, 543–548. doi:10.1016/j.fertnstert.2009.03.051

Unno, J., Takagi, M., Piao, J., Sugimoto, M., Honda, F., Maeda, D., Masutani, M., Kiyono, T., Watanabe, F., Morio, T., Teraoka, H., Mizutani, S., 2013. Artemis-dependent DNA double-strand break formation at stalled replication forks. Cancer Sci. 104, 703–710. doi:10.1111/cas.12144

Vijayraghavan, S., Schwacha, A., 2012. The eukaryotic Mcm2-7 replicative helicase. Subcell. Biochem. 62, 113–134. doi:10.1007/978-94-007-4572-8_7

Wang, F., Tian, X., Zhou, Y., Tan, D., Zhu, S., Dai, Y., Liu, G., 2014. Melatonin Improves the Quality of In Vitro Produced (IVP) Bovine Embryos: Implications for Blastocyst Development, Cryotolerance, and Modifications of Relevant Gene Expression. PLOS ONE 9, e93641. doi:10.1371/journal.pone.0093641

Watson, J.D., Crick, F.H.C., 1953. Genetic implications of the structure of deoxyribonucleic acid. Nature 171, 964–967.

Westphal, L.M., Hinckley, M.D., Behr, B., Milki, A.A., 2003. Effect of ICSI on Subsequent Blastocyst Development and Pregnancy Rates. J. Assist. Reprod. Genet. 20, 113–116. doi:10.1023/A:1022678807398

Wong, P.G., Winter, S.L., Zaika, E., Cao, T.V., Oguz, U., Koomen, J.M., Hamlin, J.L., Alexandrow, M.G., 2011. Cdc45 limits replicon usage from a low density of preRCs in mammalian cells. PloS One 6, e17533. doi:10.1371/journal.pone.0017533

Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E.J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., McDonald, S., McLean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'Neil, S., Pearson, D., Quail, M.A., Rabbinowitsch, E., Rutherford, K., Rutter, S., Saunders, D., Seeger, K.,

Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R.G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Düsterhöft, A., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T.M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dréano, S., Gloux, S., Lelaure, V., Mottier, S., Galibert, F., Aves, S.J., Xiang, Z., Hunt, C., Moore, K., Hurst, S.M., Lucas, M., Rochet, M., Gaillardin, C., Tallada, V.A., Garzon, A., Thode, G., Daga, R.R., Cruzado, L., Jimenez, J., Sánchez, M., del Rey, F., Benito, J., Domínguez, A., Revuelta, J.L., Moreno, S., Armstrong, J., Forsburg, S.L., Cerutti, L., Lowe, T., McCombie, W.R., Paulsen, I., Potashkin, J., Shpakovski, G.V., Ussery, D., Barrell, B.G., Nurse, P., Cerrutti, L., 2002. The genome sequence of Schizosaccharomyces pombe. Nature 415, 871–880. doi:10.1038/nature724

Woodward, A.M., Göhler, T., Luciani, M.G., Oehlmann, M., Ge, X., Gartner, A., Jackson, D.A., Blow, J.J., 2006. Excess Mcm2-7 license dormant origins of replication that can be used under conditions of replicative stress. J. Cell Biol. 173, 673–683. doi:10.1083/jcb.200602108

Xu, W., Aparicio, J.G., Aparicio, O.M., Tavaré, S., 2006. Genome-wide mapping of ORC and Mcm2p binding sites on tiling arrays and identification of essential ARS consensus sequences in S. cerevisiae. BMC Genomics 7, 276. doi:10.1186/1471-2164-7-276

Yang, S.C.-H., Gauthier, M., Bechhoefer, J., 2009. Computational Methods to Study Kinetics of DNA Replication, in: Vengrova, S., Dalgaard, J.Z. (Eds.), DNA Replication, Methods in Molecular Biology. Humana Press, pp. 555–573.

Yates, R.D., Goodman, D.J., 2004. Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers, 2nd Edition edition. ed. John Wiley & Sons, Hoboken, NJ.

Zeman, M.K., Cimprich, K.A., 2014. Causes and consequences of replication stress. Nat. Cell Biol. 16, 2–9. doi:10.1038/ncb2897

Zheng, L., Shen, B., 2011. Okazaki fragment maturation: nucleases take centre stage. J. Mol. Cell Biol. 3, 23–30. doi:10.1093/jmcb/mjq048

Zhu, Y.O., Siegal, M.L., Hall, D.W., Petrov, D.A., 2014. Precise estimates of mutation rate and spectrum in yeast. Proc. Natl. Acad. Sci. 111, E2310–E2318. doi:10.1073/pnas.1323011111